

DTIC FILE COPY

(2)

AD-A202 461

Research Laboratory of Electronics  
Massachusetts Institute of Technology

## Speech Database Development

Final Technical Report  
for Contract #N00039-85-C-0341

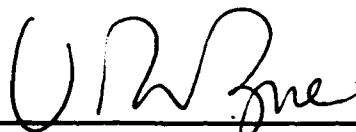
June 27, 1985 - June 26, 1987  
submitted to the  
Defense Advanced Research Projects Agency

DTIC  
ELECTE  
DEC 07 1988

S

D

CS D



Victor W. Zue  
Principal Research Scientist  
Department of Electrical Engineering  
& Computer Science  
Telephone: (617) 253-8513

**DISTRIBUTION STATEMENT A**

Approved for public release;  
Distribution Unlimited

November 21, 1988

88 12 5 074

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED			1b. RESTRICTIVE MARKINGS		
2a. SECURITY CLASSIFICATION AUTHORITY			3. DISTRIBUTION / AVAILABILITY OF REPORT		
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE			Approved for public release; distribution unlimited		
4. PERFORMING ORGANIZATION REPORT NUMBER(S)			5. MONITORING ORGANIZATION REPORT NUMBER(S)		
6a. NAME OF PERFORMING ORGANIZATION Research Laboratory of Electronics Massachusetts Institute of Technology		6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION Department of the Navy Naval Electronics Systems Command		
6c. ADDRESS (City, State, and ZIP Code) 77 Massachusetts Avenue Cambridge, MA 02139		7b. ADDRESS (City, State, and ZIP Code) Washington, DC 20363			
8a. NAME OF FUNDING / SPONSORING ORGANIZATION Advanced Res. Projects Agency		8b. OFFICE SYMBOL (If applicable) No. 5451	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N00039-85-C-0341		
8c. ADDRESS (City, State, and ZIP Code)		10. SOURCE OF FUNDING NUMBERS			
		PROGRAM ELEMENT NO.	PROJECT NO. DX-062	TASK NO.	WORK UNIT ACCESSION NO.
11. TITLE (Include Security Classification) Speech Database Development. Final Technical Report.					
12. PERSONAL AUTHOR(S) Victor W. Zue					
13a. TYPE OF REPORT Final Report		13b. TIME COVERED FROM 6/27/85 TO 6/26/88		14. DATE OF REPORT (Year, Month, Day) 11/21/88	
15. PAGE COUNT 42 pp.					
16. SUPPLEMENTARY NOTATION					
17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)		
FIELD	GROUP	SUB-GROUP			
19. ABSTRACT (Continue on reverse if necessary and identify by block number)					
Work by Victor W. Zue and his collaborators is summarized here.					
20. DISTRIBUTION / AVAILABILITY OF ABSTRACT <input checked="" type="checkbox"/> UNCLASSIFIED/UNLIMITED <input type="checkbox"/> SAME AS RPT <input type="checkbox"/> DTIC USERS			21. ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED		
22a. NAME OF RESPONSIBLE INDIVIDUAL Barbara Passero RLE Contract Reports			22b. TELEPHONE (Include Area Code) (617) 253-2566		22c. OFFICE SYMBOL

## DISTRIBUTION LIST

Director Defense Advanced Research Project Agency 1400 Wilson Boulevard Arlington, Virginia 22209 Attn: Program Management	(3)
Space and Naval Warfare System - Command Department of the Navy Washington, D. C. 20363	(2)
Administrative Contracting Officer Room E19-628 Massachusetts Institute of Technology Cambridge, Massachusetts 02139	(1)
Director Defense Advanced Research Project Agency Attn: TIO/Admin 1400 Wilson Boulevard Arlington, Virginia 22209	(1)
Defense Technical Information Center Bldg. 5, Cameron Station Alexandria, Virginia 22314	(12)
Director Naval Research Laboratory Attn: Code 2627 Washington, D. C. 20375	(6)
TACTEC Battelle Memorial Institute 505 King Avenue Columbus Ohio 43201	(1)

# Contents

<b>1 Introduction</b>	<b>2</b>
<b>2 Task Description</b>	<b>3</b>
2.1 Corpus Design . . . . .	3
2.2 Analysis of Phonetic Coverage . . . . .	4
2.3 Data Collection . . . . .	6
2.4 Automatic Transcription Alignment System Development . . . . .	6
2.5 Transcription and Alignment . . . . .	9
2.5.1 The Acoustic-Phonetic Label Set . . . . .	9
2.5.2 Criteria for Boundary Assignments . . . . .	10
2.5.3 Procedure for Entering the Aligned Transcription . . . . .	11
2.6 Database Management System Development . . . . .	13
2.7 Database Distribution . . . . .	14
<b>3 References</b>	<b>15</b>
<b>A Database Design and Analysis</b>	<b>16</b>
<b>B Automatic Alignment System</b>	<b>27</b>
<b>C Transcription and Analysis</b>	<b>32</b>



Accession For	
NTIS CRA&I	<input checked="" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution	
Availability Codes	
Dist	Availability Codes
A-1	

# 1 Introduction

The Speech Communication Group at the Research Laboratory of Electronics, Massachusetts Institute of Technology submits the final report for contract N00039-85-C-0341, awarded by the Information Science Technology Office of the Defense Advanced Research Project Agency, as monitored by the Naval Space and Warfare Systems Command. The contract covers the 24-month period starting on June 27, 1985, and is awarded for the development of an acoustic-phonetic database, to be used by the research community of the Strategic Computing Speech Program.

*an acoustic-phonetic*  
The development of the database is thought to be crucial to the speech program because the acoustic realization of phonemes depends on complex interactions among a multitude of factors. Therefore, in order to successfully develop a speaker-independent, phonetically-based speech recognition system, a large body of speech data, collected from many speakers, is needed to help us discover and quantify these context-dependent phenomena. In addition, the speech database can serve two other functions. First, it can be used for training certain speech recognition systems. For some algorithms, such as hidden Markov modelling (HMM), a large amount of training data is needed to obtain stable estimates of the parameters of the stochastic models. For rule-based algorithms, substantial amounts of data are also needed in order to set proper thresholds on speech parameters. Second, the database can be used for performance evaluation. Given the many different approaches to the speech recognition problem, it is often difficult to compare their relative merits. Testing specific recognition algorithms or entire speech recognition systems on a common database will provide a means to evaluate their relative performance. *Known as*

*System (training) for time analysis, speech synthesis*  
The specific responsibilities of MIT in developing the database were:

- To take primary responsibility for the design of an acoustic-phonetic corpus, and to provide a detailed analysis of that corpus,
- To coordinate with researchers at Texas Instruments and elsewhere in the specification of the recording procedures for the database,
- To develop a semiautomatic system to align the transcriptions with the speech waveform, with associated capabilities for researchers to modify and correct the resulting alignments,
- To provide time-aligned orthographic and phonetic transcriptions for the recorded sentences,

- To develop a database management program, so that researchers can easily access parts of the database by specifying constraints on the phonetic or lexical environment of interest, and/or the speaker's dialect and sex, and
- To transfer the speech database to NBS, and to coordinate with researchers at NBS in specifying the procedure for distributing it.

In the next section, we describe in detail the various tasks associated with MIT's database development effort. Due to the fact that the size of the database is considerably larger than we originally proposed, the development of the database was not completed until June, 1988.

## 2 Task Description

### 2.1 Corpus Design

The database design is the result of a joint effort between MIT, SRI, and TI. The corpus is comprised of 2342 distinct sentences from three different sets:

1. Two (2) *speaker calibration sentences*, provided by SRI, designed to incorporate phonemes in contexts where significant dialectical differences are anticipated. These two sentences were spoken by all talkers.
2. Four hundred and fifty (450) *phonetically compact sentences*, hand-designed by MIT with emphasis on as complete a coverage of phonetic pairs as is practical. Each sentence was spoken by seven talkers, in order to provide a feeling for speaker variation.
3. One thousand, eight hundred and ninety (1890) *randomly selected sentences*, chosen by TI, providing alternate contexts and multiple occurrences of the same phonetic sequence in different word sequences. These were chosen primarily from the "Brown corpus" of American English sentences [1], along with a few sentences from the Hultzen et al. corpus of playwrights' dialogue.

This combination of sentences was selected for its ability to balance the conflicting desires for compact phonetic coverage, contextual diversity, and speaker variability. It was decided by the research community that these three criteria were paramount to the initial acoustic-phonetic database. Each speaker read the 2 calibration sentences, 5 of the phonetically-compact sentences, and 3 of the randomly selected sentences,

providing a total of 10 sentences.

The set of 450 compact and comprehensive sentences was developed at MIT using an iterative method [2]. Using ALEXIS [3] and the Merriam-Webster Pocket dictionary (Pocket), we interactively created sentences and analyzed the resulting corpus. We began with a corpus created for the MIT speech spectrogram reading course, which included basic phonetic coverage and varying phonetic environments. Examining pairs of phonemes we augmented these sentences, attempting to have at least one occurrence of each phoneme doublet. ALEXIS was used to search the Pocket dictionary for words having phoneme sequences that were not represented and for words beginning or ending with a specific phoneme. We then created sentences using the new words and added them to the corpus. Certain difficult sequences were emphasized, such as vowel-vowel and stop-stop sequences. For a more detailed description, the reader is referred to Appendix A.

## 2.2 Analysis of Phonetic Coverage

This section describes the phonetic coverage of the compact sentence set developed at MIT and the resulting corpus of combined MIT and TI sentences (heretofore referred to as the acoustic-phonetic database, or APDB). This analysis does not include the calibration sentences as we consider their use to be of a different nature.

Table 1 compares some of the distributional properties of the APDB with three other databases: the Merriam Webster pocket dictionary (Pocket), the Harvard Lists of phonetically-balanced sentences (HL), the MIT-selected sentences (MIT-450), and the APDB sentences. The APDB include seven copies of each MIT-450 sentence, to account for the number of talkers per sentence, and a single copy of each randomly selected sentence (TI-1890).

As the table reveals, the proportion of unique words relative to the total number of words is substantially larger in the MIT-450 than the APDB, probably due to the selection procedure. Whenever possible, new words were used in sentences and to avoid duplication. Roughly 50% of the MIT-450 words are unique, as compared to only 15% of the APDB words. The TI-1890 sentences are, on the average, slightly longer than those in the MIT-450. The 10 most frequently occurring words for all of the corpora are function words or pronouns. In both the MIT-450 and the APDB corpora, the most common word is "the," accounting for roughly 7% of all words.

Not all the words in the APDB occur in Pocket. For these cases, we generated

Table 1: Description of Databases

	POCKET	HL	MIT-450	APDB
no. sentences		720	450	5040
no. unique words	19,837	1894	1792	6103
no. words	19,837	5745	3403	41,161
ave no. words/sent		7.9	7.6	8.2
min no. words/sent		5	4	2
max no. words/sent		12	13	19
ave no. syls/word	1.38*	1.1	1.58	1.54
ave no. phones/word	3.34*	2.97	4.0	3.89

\* The ave. no. syls/word and ave. no. phones/word have been weighted by Brown Corpus[1] word frequencies.

phonemic transcriptions by rule-based expansion of the dictionary entries, or, as a last resort, by a text-to-speech synthesizer. We expect that there are pronunciation variations between the dictionary and the text-to-speech synthesizer, particularly with respect to vowel color. There may also be some pronunciation errors, but we think these will be statistically insignificant.

Table 2 shows the distribution of within-word consonant sequences for the four databases. The APDB has more complete coverage of consonant sequences than the MIT-450, particularly for the word-final and word-medial sequences. We examined a list of all of the word-initial and word-final clusters in the sentence list, and compared these with the occurrences in Pocket. We verified that essentially every initial cluster that occurred more than once in the Pocket lexicon was included at least once in the APDB, and that most of the final clusters were covered. Often, if a word-final cluster did not occur in word-final position in the APDB, the sequence did occur within a word or across a word boundary. Generally, the sequences occurring in Pocket that are not covered by APDB are from borrowed words such as "moire" and "svelte." The APDB also contains many word-final consonant sequences that were not present in MIT-450. Since the Pocket lexicon does not include suffixes, there are more word-final consonant sequences in the APDB. The reader is referred to Appendix A for tables summarizing further analyses of the properties of the APDB.



Table 2: Distribution of Consonant Sequences

	POCKET	HL	MIT-450	APDB
no. unique words	19,837	1894	1792	6103
no. word-initial	75	59	64	68
no. word-final	129	105	102	146
no. word-medial	608	123	228	388

## 2.3 Data Collection

The recording of the data was primarily TI's responsibility, although we provided limited help in specifying such things as the recording environment, the types of microphone, and the sampling rate. Speech data was collected and recorded utilizing the Vocabulary Master Library file (VML). 630 VML files were created and run on the STERIODS system (VAX Fortran automated speech data collection system, also known as the STEReO automatic Interactive Data collection System), developed by TI [4].

The speech data was digitally recorded at 20 kHz in a relatively quiet environment, simultaneously on a pressure-sensitive microphone and on a Sennheiser close-talking microphone. Digital tapes were shipped to NBS, where they were filtered and downsampled to 16 kHz. The speech data was then sent to MIT to generate the orthographic and phonetic transcriptions. (The transcriptions procedures are described in the next section and in Appendices B and C.)

Each of 630 speakers, from 8 dialectal regions of the United States, read 10 sentences. Table 3 provides a summary of the number of speakers from each of the dialect regions. Approximately 70% of the speakers (439) are male and 30% are female.

## 2.4 Automatic Transcription Alignment System Development

The large amount of acoustic data constitutes only one part of the speech database. The utterances must also be augmented with a set of time-aligned transcriptions which enable the user to have direct access to specific portions of the speech signal. Thus, for example, a researcher is able to query the database for all occurrences of the phoneme /t/ preceding a stressed vowel. In addition, the researcher has the ability

Table 3: Distribution of Speakers

region	location	# speakers
1	New England	49
2	Northern	101
3	North Midland	102
4	South Midland	100
5	Southern	99
6	New York City	46
7	Western	107
8	Army Brat <sup>1</sup>	33
Total:		630

<sup>1</sup> The term "Army Brat" is used to denote speakers who lived in several geographical areas during their early lives.

to pinpoint the locations of the consonantal closures and releases and to make measurements based on the time-aligned transcription.

Traditionally, the alignment of a phonetic transcription with the corresponding speech waveform is done manually by a trained acoustic-phonetician. This is an extremely time-consuming procedure, requiring the expertise of one of a very small number of people. Therefore, the amount of data that can be labeled is limited. Manual labeling often involves decisions that are highly subjective - yielding a lack of consistency and reproducibility of results so that the results can vary substantially from one person to another. In the past few years, several automatic time-alignment procedures have been suggested [5,6,7,8,9]. The general approach is to align the acoustic waveform with a "reference" waveform, using dynamic programming algorithms. The reference is either a waveform generated from a phonetic transcription by synthesis techniques, or by using a previously labeled utterance having the same transcription. However, this approach requires either that the synthesis technique be of high quality or that the two utterances have identical phonetic transcriptions (which is rare across speakers).

Transcription alignment of the TIMIT database utilizes CASPAR, an automatic alignment system developed at MIT. Description of preliminary implementations of CASPAR can be found elsewhere [10,11]. (One of these descriptions is attached with this report as Appendix B.) Basically, phonetic alignment is accomplished in three steps. First, each 5 ms frame of the speech data is assigned to one of five broad-

class labels: *sonorant*, *obstruent*, *voiced-consonant*, *nasal/voicebar*, and *silence*, using a non-parametric pattern classifier. The assignment process uses a binary decision tree, based on a set of acoustically motivated features. Each sequence of identically-labelled frames is then collapsed into a segment of the same label, thus establishing a broad-class segmentation of the speech. Next, the output of the initial classifier is aligned with the phonetic transcription using a search strategy with some look-ahead capability, guided by a few acoustic-phonetic rules. The resulting alignment provides "islands of reliability" for more detailed segmentation and refinement of boundaries. Further segmentation of acoustic regions that correspond to two or more phonetic events after preliminary alignment is done using specific algorithms based on knowledge of the phonetic context. In some instances, heuristic rules are invoked to assign consistent but somewhat arbitrary boundaries (as between a vowel and a semivowel).

It was discovered in a formal evaluation that CASPAR can correctly perform over 95% of the labeling task previously done by human transcribers. The boundary locations produced by the system agree well with those produced by human transcribers. For example, over 75% of the automatically generated boundaries were within 10 msec of a boundary entered by a trained phonetician.

Whenever the automatic alignment system makes a mistake in boundary location, or when it fails to find a single possible alignment, human intervention is necessary. Regardless, the output of the alignment system must be certified by an experienced acoustic-phonetician. Therefore, a set of rules was specified so that boundary locations were placed as consistently as possible from transcriber to transcriber. These rules are described in Section 2.5.2. For a more in-depth discussion of the boundary criteria, the reader is referred to Appendix C.

Since the early implementation of CASPAR, as described in the literature, two major modifications have been made. First, the second module of the system which aligns the acoustic labels with the phonetic symbols has been cast into a probabilistic framework. By using a large amount of speech data for training, a set of context-dependent and durational statistics were obtained. As a result, the system has been found to be more robust. Second, a new fourth module has been added to the system to improve the resolution of the boundaries. This module computes appropriate acoustic attributes at a high analysis rate using different window shapes that depend on the specific context. The boundaries are then adjusted based on these new attributes.

## 2.5 Transcription and Alignment

A detailed description of the procedure used for entering the aligned transcription is provided in Section 2.5.3. For further analysis on this process, the reader is referred to Appendix C.

### 2.5.1 The Acoustic-Phonetic Label Set

The label set used to provide the time-aligned acoustic-phonetic sequence is intended to represent a level somewhat intermediate between phonemic and acoustic. Our belief was that clear acoustic boundaries in the waveform should all be marked, and that the criteria for positioning the boundaries between units should in part be based on our ability to mark them consistently.

In addition to the phonemes, the set of recognized acoustic-phonetic labels includes:

- Stop closures, as stops are characterized by a sequence of a closure and a release,
- Stop allophones, the glottal stop [ʔ] and the flap [ɾ], with a separate flapping decision for /t/ and /d/,
- Two allophones of /h/: voiced [h] and unvoiced [ɦ], based on an analysis of the waveform for clear low frequency periodicity,
- Four diphthongs, /aʲ/, /ɔʲ/, /eʲ/, and /aʷ/, each represented as a single label with no separate region defined for the offglide portion,
- Two vowel phonemic forms represented by more than one allophone: schwa and /u/. For /u/ a back [u] and front [ū] allophone are recognized. Schwa has four separate allophones: back [ə], front [i], retroflex [ɤ] and devoiced [ɤ̥].
- Four syllabic consonants: [m̩, n̩, ŋ̩, l̩].

Our label set also includes a category "epenthetic silence," which we use to mark acoustically distinct regions of weak energy separating sounds that involve a change in voicing. These short gaps are typically due to articulatory timing errors. The most common occurrences of such gaps are between an /s/ and a following semivowel or nasal, as in "small" or "swift."

In general, we tried to label what we heard/saw rather than what we expected. Thus, if a person said "imput" for "input," the nasal would be marked as an /m/. However, in conditions of ambiguity, the underlying phonemic form was preferentially

selected. For further listings and definitions of the label set, the reader is referred to Appendix C.

### 2.5.2 Criteria for Boundary Assignments

Often the boundaries between two acoustic-phonetic units are clear and well-defined. However, there are a number of cases where the exact placement of a boundary is problematic, or where it is not clear whether a region should be represented as one or two acoustic-phonetic units. We tried to define a set of criteria that would be systematic and least prone to human error, in order to produce boundary positionings that were as consistent as possible.

As mentioned previously, we decided that the boundary between the closure interval and the release of a stop is an important one that should be marked. It is certainly a very distinct landmark in the waveform. Anyone interested in studying the burst characteristics of a stop would then be able to focus on just the region that includes the released portion. In a strictly phonemic representation, the closure and release would be represented as a single unit, leaving that critical boundary unmarked.

A type of problematic boundary is the one separating a prevocalic stop from a following semivowel, as in "truck." Typically, part of the /r/ is devoiced, and therefore is absorbed into the aspiration portion of the stop. If listening were the only criterion, then the left boundary of the /r/ would occur somewhere in the aspiration, and the right boundary would occur somewhere after voicing onset. A clear acoustic boundary at the point of voice onset would remain unmarked. It would also be difficult to decide where to mark the boundary between the stop burst and the aspirated /r/ portion. Since voice-onset time (VOT) is a parameter that has been a focus of many research projects, it seems unsatisfactory not to include a reliable mechanism for measuring VOT based on the labelled boundaries. Therefore, we adopted the policy of always absorbing into the stop release all of the unvoiced portion of a following semivowel.

The right-hand boundary of many prevocalic semivowels, and the left-hand boundary of post-vocalic semivowels are both rather ill-defined in the spectrogram, because the transitions are slow and continuous. It is not possible to define a single point in time that separates the vowel from the semivowel. In such cases, we decided to adopt a  $1/3 - 2/3$  formula, giving the vowel twice as much duration as the semivowel.

Whenever the same phoneme follows itself (gemination), we did not attempt to mark a boundary between the two units. This situation occurs exclusively at word

boundaries, as in "some money." Furthermore, in the case of a stop-stop sequence where the first stop is unreleased, the closure interval was identified with the first stop and the release with the second one.

### **2.5.3 Procedure for Entering the Aligned Transcription**

The labelling process involved three steps:

1. An acoustic-phonetic sequence was entered by hand as a string.
2. The speech waveform was aligned automatically with the acoustic-phonetic sequence, using the system described in Section 2.4.
3. The automatically generated boundaries were hand corrected.

In steps 1 and 3, the labeler made use of the displayed spectrogram, spectral cross sections, the original waveform, and auditory output. This process took place within the SPIRE software facility for analyzing speech, a powerful interactive tool that is well-matched to this task [3].

The first step required less intensive use of the SPIRE tool than the third step, because it was only necessary to record what was heard, without identifying the time locations of the events. The labels were entered either by typing or by mousing from a displayed set. Judgments were made using the spectrogram and waveform displays, as well as careful listening.

Once a phonetic sequence has been provided, a preliminary alignment of the phonetic symbols with the waveform is performed using CASPAR, as described in Section 2.4. Any errors in the automatically aligned acoustic-phonetic sequence is then corrected by hand. Hand-correction was based on critical listening of portions of the utterance as well as visual examination of the spectrogram and the waveform. The SPIRE layout for this stage is shown in Figure 1. As illustrated in the figure, the transcription boundaries are overlaid on the spectrogram for ease of decision-making. The spectrogram covers close to 3 seconds of speech at one time, whereas the waveform is displayed on a much more expanded time scale. Any subportions of the waveform or of the spectrogram could be used to define regions to be played out to earphones. In addition, a phonetically-labelled region in the spectrogram could be moused such that only the portion between the two boundaries was played.

The mouse was used to move existing boundaries to new points in time, to erase boundaries, or to insert new boundaries. In addition, specified mouse clicks on any

Phonetic Transcription Layout 1

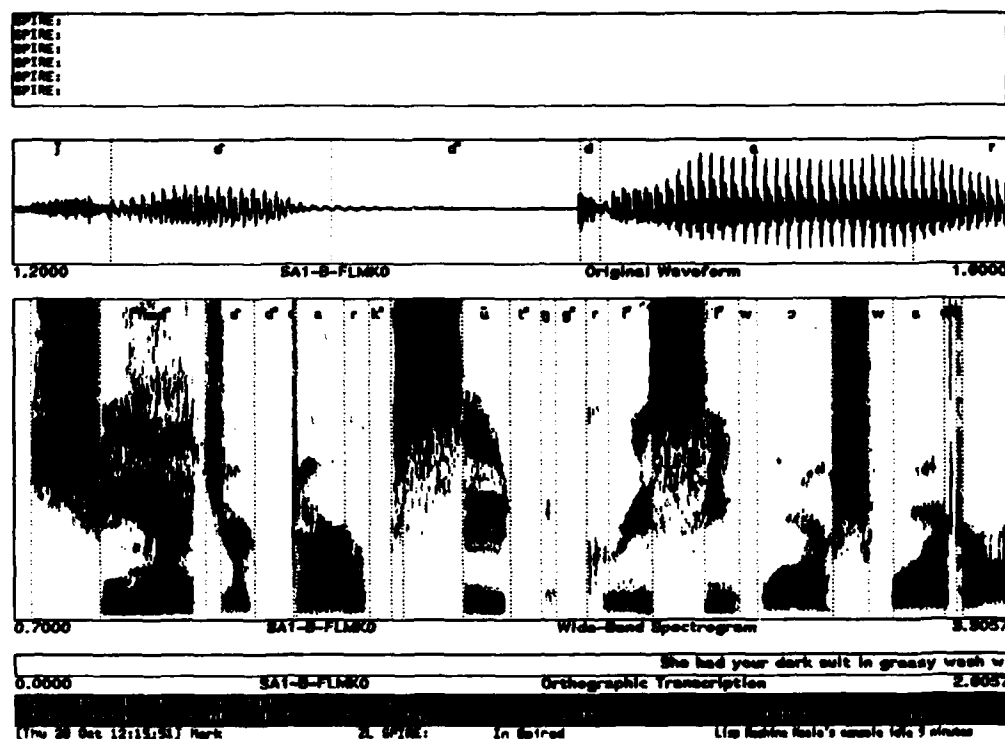


Figure 1: Alignment correction layout.

segment allowed the labeler to change the acoustic-phonetic label associated with that segment. This step was occasionally necessary to correct an error of judgment made in step 1.

Once the phonetic transcription is aligned, it is rather straightforward to propagate the alignment up to the orthographic transcription as well as the intermediate phonemic transcription. A time-aligned orthographic transcription is useful when searching for a specific word, while time-aligned phonemic transcription can be used to relate the lexical representation of words to their acoustic realizations.

In addition to the acoustic-phonetic alignment system described above, we have also developed a system that maps a time-aligned acoustic-phonetic transcription to the phonemic and orthographic transcriptions [12]. However, the alignment effort for these transcriptions lags somewhat behind the phonetic alignment. We will provide these transcriptions in a future release.

## 2.6 Database Management System Development

We have implemented an utterance database management/access system (DBMS) based on the SEARCH [3] statistical analysis tool. The SPIRE Utterance Database System (or SUDS) works by "scanning" a user-defined database of utterances. The scanning procedure is much less time- and memory-intensive than actually loading the utterances. A SEARCH sample is then built from the database, and all the power of SEARCH may be utilized to deal with the problem being examined. Utterance databases (and their associated samples) may be saved, loaded, and "rescanned" or revised. The rescanning procedure is faster than scanning, and consists of updating information about those files that have changed, and then rebuilding the sample. Rescanning is useful when, for example, a few transcriptions have been changed in a database of hundreds of utterances.

Commands already implemented in SUDS include Load Utterances, Delete Utterances, and Dump Utterances, which can be used for making tapes of the selected utterances. There is also a Save Utterance List command, which can be used to write a file with the filenames of the selected utterances. The file can then be processed by arbitrary user code. The entire system is designed for easy extensibility, and other commands can be added with very little effort.



## **2.7 Database Distribution**

The acoustic-phonetic database was completely phonetically transcribed, aligned, and checked as of June 1988. As the sentences were completed, they were sent to NBS, where they were examined and prepared for distribution. The distribution is being accomplished in three balanced stages, of which two-third has already been released. Currently, the database is available to general public via magnetic tapes, although plans for compact disc releases are well under way.

Many minor errors in the database have been found and corrected, both at MIT and at NBS, but despite our best intentions, more errors undoubtedly exist. It is our intention to continually provide corrections and updates for the foreseeable future.

### 3 References

- [1] Kucera, H. and W.N. Francis, *Computational Analysis of Present-Day American English*, Brown University Press, Providence, R.I.
- [2] Lamel, L. F., R. H. Kassel, and S. Seneff, "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus," *Proc. DARPA Speech Recognition Workshop*, Report No. SAIC-86/1546, pp. 100-109, February 1986.
- [3] Cyphers, D. S., R. H. Kassel, D. H. Kaufman, H. C. Leung, M. A. Randolph, S. Seneff, J. E. Unverferth, III, T. Wilson, and V. W. Zue, "The Development of Speech Research Tools on MIT's Lisp Machine-Based Workstations," *Proc. DARPA Speech Recognition Workshop*, Report No. SAIC-86/1546, pp. 110-115, February 1986.
- [4] Fisher, W. et al. "The DARPA Speech Recognition Research Database: Specifications and Status," *DARPA Speech Recognition Workshop Proceedings*, Texas Instruments Inc., Computer Sciences Center, February 1986.
- [5] Lowry, M.R., "Automatic Labeling of Speech From Phonetic Transcription," S.M. thesis, Massachusetts Institute of Technology, 1978.
- [6] Wagner, M., "Automatic Labelling of Continuous Speech with a Given Phonetic Transcription Using Dynamic Programming Algorithms," *ICASSP 1981*.
- [7] Chamberlain, R.M., and J.S. Bridle, "ZIP: A Dynamic Programming Algorithm for Time-aligning Two Indefinitely Long Utterances," *ICASSP 1983*.
- [8] Hohne et al., "On Temporal Alignment of Sentences of Natural and Synthetic Speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-31, No. 4, August 1983.
- [9] Lennig, M., "Automatic Alignment of Natural Speech with a Corresponding Transcription," *Speech Communication, 11th International Congress on Acoustics, Toulouse July 1983*.
- [10] Leung, H. C., and V. W. Zue, "A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech," *Proc. ICASSP-84*, pp. 2.7.1-2.7.4, March 1984.
- [11] Leung, H. C., "A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech," S.M. Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, January, 1985.
- [12] Kassel, R. H., "Aids for the Design, Acquisition, and Use of Large Speech Databases," S.B. thesis, Massachusetts Institute of Technology, May 1986.

## **A Database Design and Analysis**

Lamel, L. F., R. H. Kassel, and S. Seneff, "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus," *Proc. DARPA Speech Recognition Workshop*, Report No. SAIC-86/1546, pp. 100-109, February 1986.

# SPEECH DATABASE DEVELOPMENT: DESIGN AND ANALYSIS OF THE ACOUSTIC-PHONETIC CORPUS\*

Lori F. Lamel, Robert H. Kassel, and Stephanie Seneff

Department of Electrical Engineering and Computer Science, and  
Research Laboratory of Electronics  
Massachusetts Institute of Technology  
Cambridge, Massachusetts 02139

## ABSTRACT

The need for a comprehensive, standardised speech database is threefold: first, to acquire acoustic-phonetic knowledge for phonetic recognition; second, to provide speech for training recognisers; and third, to provide a common test base for the evaluation of recognisers. There are many factors to consider in corpus design, making it impossible to provide a complete database for all potential users. It is possible, however, to provide an acceptable database that can be extended to meet future needs. After much discussion among several sites, a consensus was reached that the initial acoustic-phonetic corpus should consist of calibration sentences, a set of phonetically compact sentences, and a large number of randomly selected sentences to provide contextual variation. The database design has been a joint effort including MIT, SRI, and TI. This paper describes MIT's role in corpus development and analyses of the phonetic coverage of the complete database. We also include a description of the phonetic transcription and alignment procedure.

## INTRODUCTION

The development of a common speech database is of primary importance for continuous speech recognition efforts. Such a database is needed in order to acquire acoustic-phonetic knowledge, develop acoustic-phonetic classification algorithms, and train and evaluate speech recognisers. The acoustic realisation of phonetic segments results from a multitude of factors, including the canonical characteristics of the phoneme, contextual dependencies, and syntactic and extralinguistic factors. A large database will make it possible to examine in detail many of these factors, with the hope of eventually understanding acoustic variability well enough to design robust speech recognisers. A complete database should include different styles of speech, such as isolated words, sentences and paragraphs read aloud, and conversational speech. The speech samples should be gathered from many speakers (at least several hundred) of varying ages, both male and female,

with a good representation of the major regional dialects of American English.

## DESIGN CONSIDERATIONS

There are many factors to consider in designing a large corpus for speech analysis. Unfortunately, some of the goals are limited by practical considerations. Ideally we would like to include multiple samples of all phonemes in all contexts, a goal that is clearly impractical for a manageable database.

At the last DARPA review meeting it was decided that an initial acoustic-phonetic database would be designed to have good phonetic coverage of American English. It was agreed that the initial acoustic-phonetic corpus would include calibration sentences (spoken by every talker), a small set of phonetically compact sentences (each spoken by several talkers) and a large number of sentences (each to be spoken by a single talker). This combination was chosen to balance the conflicting desires for compact phonetic coverage, contextual diversity, and speaker variability. Another requirement of the corpus was that the sentences should be reasonably short and easy to say.

The database design is a joint effort between MIT, SRI, and TI. The speaker calibration sentences, provided by SRI, were designed to incorporate phonemes in contexts where significant dialectal differences are anticipated. They will be spoken by all talkers. The second set of sentences, the *phonetically compact* sentences, was hand-designed by MIT with emphasis on as complete a coverage of phonetic pairs as is practical. Each of these sentences will be spoken by several talkers, in order to provide a feeling for speaker variation. Since it is extremely time-consuming and difficult to create sentences that are both phonetically compact and complete, a third set of *randomly selected* sentences, chosen by TI, provides alternate contexts and multiple occurrences of the same phonetic sequence in different word sequences.

A breakdown of the actual sentence corpus is shown in Table 1. This arrangement was chosen to balance the conflicting desires for capturing inter-speaker variability and providing contextual diversity. Since the calibration

\*This research was supported by DARPA under contract N00039-85-C-0341, monitored through Naval Electronic Systems Command.

	No. Talkers	No. Sentences	Total
Calibration (SRI)	640	2	1280
Compact (MIT)	7	450	3150
Random (TI)	1	1890	1890
Total	—	—	6320

Table 1: Breakdown of Frequencies of Occurrence of Sentences in Corpus

sentences are spoken by all of the speakers, they should be useful for defining dialectal differences. For multiple instances of the exact same phonetic environments, but with a much richer acoustic-phonetic content than in the calibration sentences, the MIT set would be appropriate. The TI sentences, to be spoken by one talker per sentence, should provide data for phoneme sequences not covered by the MIT database.

## DESIGN OF THE COMPACT ACOUSTIC-PHONETIC SENTENCES

A set of 450 sentences was hand-designed at MIT, using an iterative procedure, to be both compact and comprehensive. We made no attempt to phonetically balance the sentences. We used *ALexis* and the Merriam-Webster Pocket Dictionary (Pocket) to interactively create sentences and analyse the resulting corpus. We began with the "summer" corpus created for the MIT speech spectrogram reading course to include basic phonetic coverage and interesting phonetic environments. We initially augmented these sentences by looking at pairs of phonemes, trying to have at least one occurrence of each phoneme pair sequence. *ALexis* was used to search the Pocket dictionary for words having sequences that were not represented and for words beginning or ending with a specific phoneme. We then created sentences using the new words and added them to the corpus. Certain difficult sequences were emphasised, such as vowel-vowel and stop-stop sequences. Some phoneme pairs are impossible; others are extremely rare and occur only across word boundaries. For example, /w/ and /y/ never close a syllable, except as an off-glide to a vowel, so many /w/-phoneme pairs are impossible. After filling some of the gaps in coverage, we reanalysed the sentences with regard to phoneme pair coverage, consonant sequence coverage, and the potential for applying phonological rules both within words and across word boundaries. In a final pass through the sentence set, we modified and enriched sentences where simple substitutions could introduce variety or generate an instance of a rare phoneme pair.

## ANALYSIS OF PHONETIC COVERAGE

This section discusses the phonetic coverage of the compact sentence set developed at MIT and the resulting cor-

pus consisting of the combined MIT and TI sentences. This analysis does not include the calibration sentences as we consider their use to be of a different nature.

	POCKET	HL	MIT-450	APDB
# sentences		720	450	5040
# unique words	19,837	1894	1792	5107
# words	19,837	5745	3403	41,161
ave # words/sent		7.9	7.8	8.2
min # words/sent		5	4	2
max # words/sent		12	13	19
ave # syls/word	1.38*	1.1	1.58	1.54
ave # phones/word	3.34*	2.97	4.0	3.89

\* The ave # syls/word and ave # phones/word have been weighted by Brown Corpus[1] word frequencies.

Table 2: Description of Databases

Table 2 compares some of the distributional properties of the Pocket Lexicon (Pocket), the Harvard List (HL)[2], the MIT-selected sentences (MIT-450), and the Acoustic-Phonetic Database selected sentences (APDB). The APDB includes seven copies of each MIT-450 sentence, to account for the number of talkers per sentence, and a single copy of each randomly selected sentence (TI-1890). Since we were given only the orthographies for the TI-1890 sentences, we generated phonemic transcriptions by dictionary lookup, by rule-based expansion of the dictionary entries, and, as a last resort, by a text-to-speech synthesiser. We expect that there are pronunciation variations between the dictionary and the text-to-speech synthesiser, particularly with respect to vowel color. There may also be some pronunciation errors, but we think these will be statistically insignificant.

The proportion of unique words relative to the total number of words is substantially larger in the MIT-450 than the APDB, probably due to the selection procedure. We tried to use new words in sentences and to avoid duplication when at all possible. Roughly 50% of the MIT-450 words are unique, as compared to only 25% of the APDB words. The TI-1890 sentences are, on the average, slightly longer than those in the MIT-450. The 10 most frequently occurring words for all of the corpora are function words or pronouns. In both the MIT-450 and the APDB corpora, the most common word is "the," accounting for roughly 7% of all words.

The average numbers of syllables and phones per word are longer for the MIT-450 and the APDB than for the HL. This is presumably due to the higher percentage of polysyllabic words.

Figure 1 shows the distribution of the number of syllables per word for the two corpora. The distributions are quite similar, with the majority of the words being mono- or bi-syllabic. The MIT-450 corpus has a slightly higher percentage of polysyllabic words than does the combined

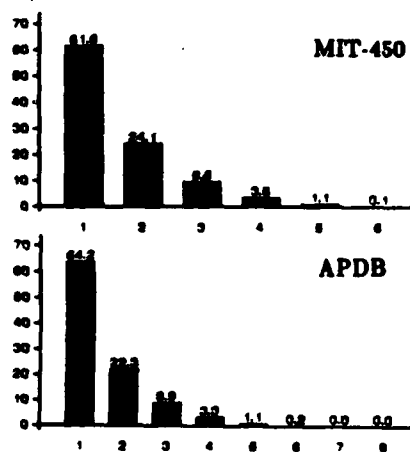


Figure 1: Histograms of the number of syllables per word.

corpus. We specifically tried to include polysyllabic words in the sentences, since these are likely to be spoken with greater variability.

Distributions of the number of phonemes per word are shown in Figure 2. The 10 most common phonemes and their frequency of occurrence are given in Figure 3.

Table 3 shows the distribution of within-word consonant sequences for the four databases. The MIT-450 sentence set covers most of the consonant sequences occurring within words. The APDB has more complete coverage, particularly for the word-final and word-medial sequences. We examined a list of all of the word-initial and word-final clusters in the sentence list, and compared these with the occurrences in Pocket. We verified that essentially every initial cluster that occurred more than once in the Pocket lexicon was included at least once in the APDB, and that most of the final clusters were covered. Often, if a word-final cluster did not occur in word-final position in the APDB, the sequence did occur within a word or across a word boundary. Generally, the sequences occurring in Pocket that are not covered by APDB are from borrowed words such "moire" and "svelte."

The APDB includes many word-final consonant sequen-

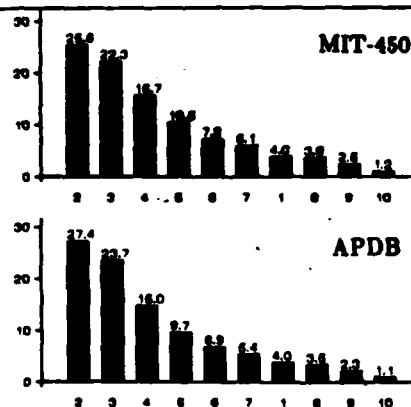


Figure 2: Histograms of the number of phonemes per word.

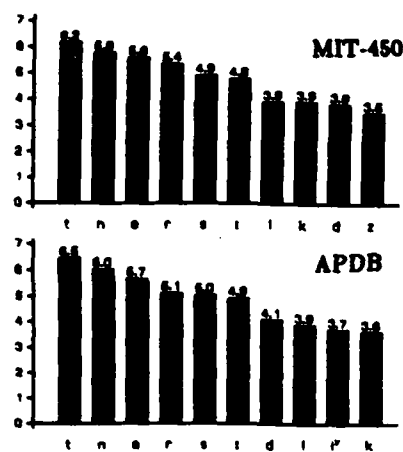


Figure 3: Histograms of the 10 most common phonemes.

	POCKET	HL	MIT-450	APDB
# unique words	19,837	1894	1792	6103
# WI	75	59	64	68
# WF	129	106	102	146
# WM	608	123	228	388
# boundaries		4305	2953	36,121
# WB		976	805	1630

Table 3: Distribution of Consonant Sequences

ces that were not present in MIT-450. In fact, there are more word-final consonant sequences in the APDB than actually occur in Pocket. The reason is that the Pocket lexicon does not include suffixes.

A more detailed phonetic analysis of all phoneme pairs is included in Appendix 1 in tabular form. The tables are broken down into phoneme subsets, and data are included for both the MIT-450 and the APDB. Some of the gaps in the MIT-450 table have been filled in by sentences in the TI-1890 corpus (e.g., the syllabic /l/ column of the vowel-sonorant pairs table and the /y/ column of the vowel-sonorant pairs table). Note also that some gaps occur in both tables. Such gaps are expected, since some phoneme sequences are impossible or quite rare. For example, the lax vowels (excluding schwa) are never found in syllable-final position in English. As a consequence, table entries requiring lax vowels as the first member of a pair have many gaps (see for example, the vowel-vowel entries in the pair tables.)

Figure 4 compares histograms of the sentence types for the MIT-450 and the APDB. Simple sentences (Simple S.) and questions (Simple Q.) have no major syntactic markers. Complex sentences (Complex S.) and questions (Complex Q.) are expected to have a major syntactic boundary when read. As can be seen, the APDB has a wider variety of sentence types, with 75% being simple declarative sentences. In the MIT-450, almost 85% of the sentences are of the simple declarative form. Questions form about

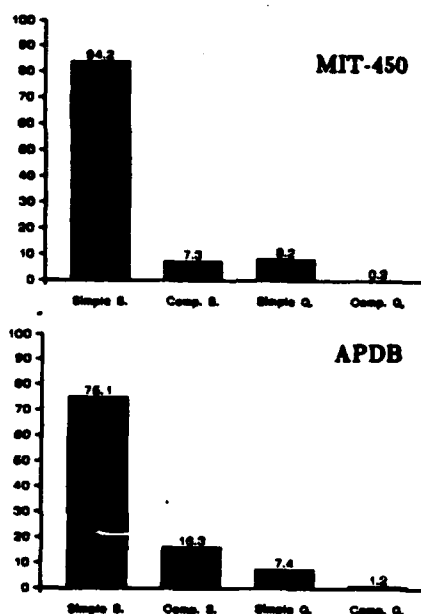


Figure 4: Histogram of sentence types.

10% of both corpora.

Figure 5 shows counts of environments where major phonological rules may apply. We chose to gather information on the following possibilities:

- gemination (GEM)
- vowel-vowel sequences (VVS)

- vowel-schwa sequences (VSS)
- schwa-vowel sequences (SVS)
- flapping of /t/, /d/, and /n/ (FLAP)
- homorganic stop insertion (HSI)
- schwa devoicing (S-DVC)
- fricative devoicing (F-DVC)
- /s/-/ʃ/ and /z/-/ʒ/ palatalisation (PAL)
- y-palatalisation: /dy/ → /j/ (DY-Jh)
- y-palatalisation: /ty/ → /tʃ/ (TY-Ch)
- y-palatalisation: /sy/ → /ʃ/ (SY-Sh)

The histograms show that both corpora have many potential environments for flapping and homorganic stop insertion. The vowel-vowel environments are also well covered. The analysis for phonological rule application is difficult, because of the difficulties in predicting what different speakers will say.

## RECORDING, LABELING, AND ALIGNMENT

The recording of the sentences is currently under way at TL. Speech is recorded digitally at 20 kHz, simultaneously on a pressure-sensitive microphone and on a Sennheiser close-talking microphone. Digital tapes are shipped to NBS, where they are filtered and downsampled to 16 kHz. The resampled tapes are then shipped to MIT where the orthographic and phonetic transcriptions are generated.

Transcriptions are generated using the *Spire* facility, in conjunction with the automatic alignment system pro-

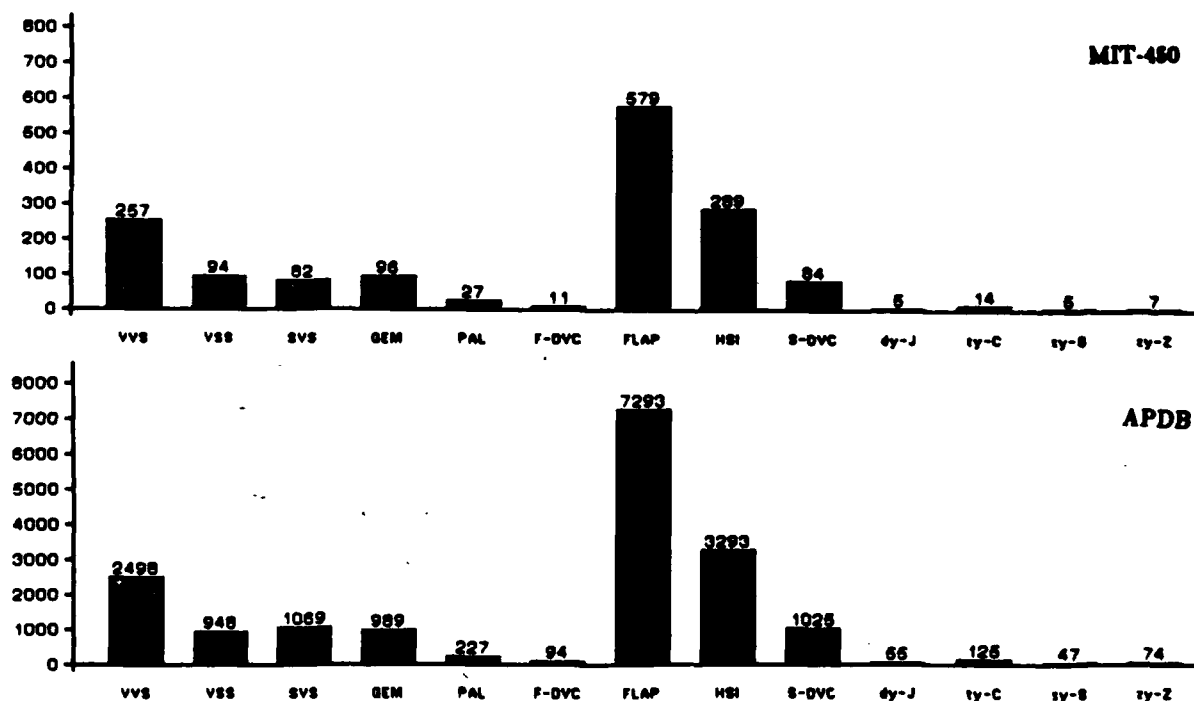


Figure 5: Histogram for potential application of phonological rules.

Unvoiced Stops:	p t k ʈ
Voiced Stops:	b d g ɟ
Stop Gaps:	ɸ ʔ ʕ ʁ ʙ ɗ ɠ ɣ
Nasals:	m n ŋ ɲ
Syllabic Nasals:	ɱ ɳ ɰ ɶ
Unvoiced Fricatives:	s ʃ f θ
Voiced Fricatives:	z ʒ v ð
Glides:	l r w y
Vowels:	ɪ ɛ ʊ ɔ ʌ ɑ ɐ ɤ
Schwa:	ə ɘ ɚ
H, Silences:	h ɦ ɔ

Figure 6: Phones used for labeling.

vided by Leung [3]. The transcription process involves three steps:

1. A "Phonetic Sequence," which consists of a list of the phones of the utterance in correct temporal order but with no boundaries marked in time, is entered.
2. The utterance is run through an automatic system to generate an alignment for the sequence.
3. The automatically generated alignment is hand-corrected.

Only the data recorded through the pressure microphone are transcribed. Transcriptions for the close-talking version are generated by duplicating the results for the pressure microphone.

The phones used in the labeling are shown in Figure 6. In many cases, it is not possible to define a boundary between two phones, such as /ɔr/, because features appropriate for both phones often occur simultaneously in time. When no obvious positioning of the boundary is apparent, arbitrary rules, such as an automatic 2/3:1/3 split, are invoked. There are also some cases in which none of our standard phones are appropriate for a given portion of the speech, primarily because of severe coarticulation effects. In such cases, the segment is labeled as the nearest phone equivalent, according to the transcriber's judgment. There are other difficult cases, such as syllable-initial /pl/, where the /l/ is devoiced at onset. Should the portion before voicing begins be thought of as part of the aspiration of the /p/, or as part of the /l/? We have decided, somewhat arbitrarily, to define the onset time of the phone following an unvoiced stop as coincident with the onset of voicing. These remarks serve simply as examples of some of the difficulties that arise in transcribing continuous speech. We are mainly interested in using consistent methods of transcribing in situations where ambiguity exists. Currently the transcription rate is 100 sentences per week.

## SUMMARY

We have described various components of the preliminary acoustic-phonetic database and discussed some of the issues in its design. Evaluating the phonetic coverage of the database is difficult primarily because no

dard for comparison exists. We have chosen to compare the phonetic coverage of the database to two well-known sources, the Merriam-Webster Pocket Dictionary of 1964 and the Harvard List sentences. The dictionary does not reflect spoken English very well, and can only guide us in judging the possible phonemic sequences within words. The Harvard List sentences, while phonemically balanced, consist primarily of very simplistic sentences and monosyllabic words. In addition, they are balanced for phoneme occurrences, whereas we tried to account for occurrences of phoneme pairs.

We believe that we have adequate coverage of most phonemes and phoneme pairs. In cases where the phoneme pairs are scarce, there are often other phoneme pairs that will provide similar information. For example, the class sequence [alveolar consonant] [back vowel] is more general than /t/ /ɔ/, and has a higher frequency of occurrence.

We hope that the APDB database will provide guidelines for the development of future databases. An analysis of the spoken corpus will enable us to judge our phonetic analysis procedure. In particular, we will be able to evaluate the relationship between our phonological rule predictions and the frequency with which a phonological modification actually occurred.

## REFERENCES

- [1] Kucera, H. and W.N. Francis (1967) *Computational Analysis of Present-Day American English*, Brown University Press, Providence, R.I.
- [2] Egan, J. (1944) "Articulation testing methods II," OSRD Report No. 3802, U.S. Dept. of Commerce Report PB 22848, November.
- [3] Leung, H. C. and V.W. Zue (1984) "A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech," *Proc. ICASSP-84*, 2.7.1-2.7.4.



# Appendix 1

## MIT

	i'	e'	ɔ	o'	u	a'	ɔ'	a''	ɜ
i'	3	7	8	9		3	1	3	1
e'	1	1	4	4		1		1	1
ɔ	1	1	1	1				1	
o'	2	2	3	1		1	1	3	
u	1	5	5	3	1	2	1	2	1
a'	4	4	6	2	1	3	1		1
ɔ'	1	1	1	1					
a''	1	2	1	1		1	1		
ɜ	7	1	1	2					1
i									
ɛ									
æ									
a	1								
ʌ									
u									
ɜ	4	7	4	1		6	1	3	1
ɔ	3	1		7		4	1	1	1
i	1								

## APDB

	i'	e'	ɔ	o'	u	a'	ɔ'	a''	ɜ
i'	28	70	75	88		27	7	26	22
e'	10	7	38	31		10		7	9
ɔ	7	7	7	7				7	
o'	14	14	26	10		9	7	21	4
u	12	48	43	26	7	19	7	16	10
a'	28	30	47	16	8	21	7	2	8
ɔ'	10	7	7	7					
a''	7	14	9	7		7	7		
ɜ	62	10	12	18		3		2	7
i									
ɛ				2					3
æ									
a	8								
ʌ									
u									
ɜ	64	73	37	13		52	10	25	7
ɔ	22	8	7	60		33	7	11	10
i	7								

	i'	e'	ɔ	o'	u	a'	ɔ'	a''	ɜ
n	38	5	8	11	10	14	4	3	
m	17	24	10	18	9	24	2	1	4
ŋ	1	1	3					1	1
ɱ			1						
ɒ		1							
l	2		2			1		1	
l	88	23	13	25	9	22	4	12	7
r	81	28	16	28	24	24	3	5	
w	20	20	20	6	1	17	1	1	10
y			1	1	88				1

	i'	e'	ɔ	o'	u	a'	ɔ'	a''	ɜ
n	422	79	88	252	117	133	32	98	21
m	246	318	100	227	88	248	18	22	48
ŋ	9	7	32	8		1	1	10	7
ɱ			7						
ɒ		7							
l	55	5	17	5		13		7	2
l	1001	287	168	286	115	308	34	110	60
r	936	312	169	303	260	272	29	71	6
w	294	212	199	67	8	210	7	9	134
y	3		9	7	933			5	10

	i'	e'	ɔ	o'	u	a'	ɔ'	a''	ɜ
b	27	2	9	10	1	28	4	3	8
d	18	22	8	10	14	7		9	2
q	1	3	1	13	4	1		1	1
p	14	14	2	11	2	9	8	3	12
t	44	19	11	13	86	18	1		4
k	9	9	19	15	7	3	4	8	11
ç	6	9	1	2	4	3	1		3
j	7	2	3	1	6	2	5		4
s	22	5	12	5	9	13	2	2	10
z	14	7	10	3	2	2		1	4
ʃ	11	2	5	5	3	1		1	1
ʒ					1				
f	9	5	17	46	2	8		3	12
v	10	6	2	5		3	2	1	4
θ	6		2					3	5
ð	3	9		5					
h	8		7	4	5	10		12	18

	i'	e'	ɔ	o'	u	a'	ɔ'	a''	ɜ
b	364	35	78	118	15	282	48	54	79
d	270	212	95	133	165	80	1	107	44
q	7	51	18	137	29	15	1	7	28
p	149	141	32	142	21	81	75	38	137
t	501	234	157	175	1060	220	9	21	99
k	99	137	191	146	56	31	32	75	106
ç	54	73	10	18	56	28	9	1	26
j	61	16	25	12	55	17	40		36
s	275	83	137	176	85	181	20	26	123
z	130	63	109	58	18	33		12	37
ʃ	198	33	42	54	35	9		11	15
ʒ					22				
f	120	72	172	506	25	117	4	32	120
v	92	68	28	44	60	28	15	63	
θ	56	2	31	3	3	3		37	53
ð	65	192		59				2	3
h	381	18	58	79	54	117		137	213

## MIT

	I	E	æ	ɑ	Λ	U	ʔ	ə	i
ɪ	11	4	16	11			5	28	1
e	4	3	3	1	1		1	3	
ɔ	1	1	1	2					1
o	4	1	2	3			1		3
u	6	7	4	6	1		1	4	
ɑ	5	2	7	3	2		4	11	6
ʊ	3	1	1				2	1	2
ɑ		1	1		1		3	3	
ɜ	2		1		1			1	1
ɪ									
ɛ									
æ									
ɑ									
Λ									
U									
ʔ	6	4	7	6	2		1	7	4
ə	5	6	8	2	2			3	
i									

## APDB

	I	E	æ	ɑ	Λ	U	ʔ	ə	i
ɪ	174	47	182	102	7		46	399	27
e	38	22	26	21	9		8	38	7
ɔ	9	7	9	14	2			3	8
o	54	14	26	25	3		28	28	34
u	75	58	48	54	11		8	79	13
ɑ	51	18	61	23	14		30	107	61
ʊ	21	8	8				14	7	15
ɑ	5	17	9	1	8		32	30	
ɜ	27	8	13	1	9			30	16
ɪ									
ɛ								1	
æ									
ɑ									
Λ									
U									
ʔ	80	38	73	48	21		8	106	52
ə	57	52	82	19	23			42	2
i									

	I	E	æ	ɑ	Λ	U	ʔ	ə	i
n	38	16	13	10	10		16	29	14
m	17	15	13	7	23		9	44	8
ŋ	12		3	1				2	2
ɪ									
ɑ	2	3	1	1				1	1
ɪ	10	2	2	1				7	3
l	37	22	29	20	11	5	12	33	24
r	69	39	33	26	38	1	2	49	16
w	60	25	3	7	7	12	6	30	
y	4	4		3	6	38	2	1	1

	I	E	æ	ɑ	Λ	U	ʔ	ə	i
n	434	185	185	260	148	3	144	388	214
m	231	223	219	93	240		91	516	87
ŋ	125	6	51	10	11			43	14
ɪ	1		3					1	
ɑ	18	24	11	9			2	11	10
ɪ	93	29	37	10	2			96	43
l	397	281	330	191	127	67	119	342	279
r	808	415	364	271	393	10	17	562	211
w	659	293	31	125	150	145	144	376	5
y	44	55	3	29	46	353	44	15	9

	I	E	æ	ɑ	Λ	U	ʔ	ə	i
b	25	12	16	15	11	3	5	12	5
d	56	9	17	6	7	1	16	56	22
q	9	10	9	13	7	4	9	3	5
p	12	13	14	20	2	5	23	16	3
t	63	21	30	16	8	2	44	47	45
k	10	11	42	25	23	6	6	33	14
ç	6	4	2	7	1		10	3	6
j	12	12	4	4	6		5	2	25
s	44	15	11	8	19	1	7	39	20
z	37	7	17	13	4		4	36	9
ʃ	5	5	3	2		9	1	3	42
ʒ						1	4		3
ɪ	23	5	6	5	5	2	2	15	2
v	12	12	10	3			33	16	3
θ	14	1	1	1			4	3	
ð	15	14	21		1		12	219	
h	24	7	31	15	3	3			

	I	E	æ	ɑ	Λ	U	ʔ	ə	i
b	271	123	185	145	204	31	69	185	41
d	646	130	209	82	121	10	174	595	245
q	99	132	97	118	67	57	87	30	54
p	169	150	166	208	41	65	222	223	44
t	764	288	326	160	105	35	456	690	481
k	127	97	415	266	251	81	51	404	178
ç	61	41	29	56	14	4	101	47	70
j	110	120	35	39	68	1	50	29	230
s	523	349	166	88	268	8	86	488	270
z	386	69	228	157	52		36	437	112
ʃ	56	48	40	32	14	112	13	84	472
ʒ			1			7	39	3	34
ɪ	268	87	96	74	46	26	39	172	31
v	152	145	124	40	11		345	230	52
θ	171	16	18	7	3		34	50	19
ð	238	196	323		14		187	2230	
h	455	118	452	137	46	27	9	2	4

## MIT

	b	d	g	p	t	k	č	ǰ
iʔ	7	27	9	27	32	25	17	1
eʔ	8	13	2	9	47	17	2	7
ɔ	2	5	5	1	13	4	3	1
oʷ	5	2	2	9	15	12	3	
u	10	14	5	15	17	10	4	7
aʔ	7	17	2	2	24	14	1	1
ɔʔ	2	22	1	2				
aʷ		2	1	1	22	1		
ʔ	8	5	1	5	10	10	6	5
i	5	13	28	16	21	87	8	12
ɛ	2	16	8	5	20	37	1	8
æ	10	10	10	9	37	27	4	2
a	13	7	3	19	31	19	2	10
ʌ	10	3	5	15	12	4	8	3
u	1	24	1	1	4	13	1	
ə	6	14	2	3	10	8	4	
ə	48	46	28	45	43	46	2	14
i	8	13	2	3	27	9		8

## APDB

	b	d	g	p	t	k	č	ǰ
iʔ	116	357	102	292	387	270	155	13
eʔ	113	188	20	84	503	228	18	62
ɔ	16	41	48	8	130	52	23	7
oʷ	68	61	30	104	157	124	27	
u	121	173	58	152	196	120	28	57
aʔ	60	249	19	31	292	166	7	8
ɔʔ	15	22	7	14	1	3		
aʷ	3	42	9	8	262	8		3
ʔ	69	75	11	53	121	112	52	46
i	48	186	266	186	501	873	100	145
ɛ	16	224	76	83	269	438	8	71
æ	95	192	99	118	519	357	43	26
a	133	91	33	195	408	183	16	87
ʌ	101	54	60	171	202	64	104	23
u	8	314	10	8	48	149	7	
ə	63	203	23	43	101	83	34	5
ə	537	616	307	477	517	514	31	116
i	79	168	28	37	387	137		75

	b	d	g	p	t	k	č	ǰ
n	13	119	7	11	120	17	5	16
m	15	3	3	33	6	6	1	1
ŋ	3	4	9	3	4	10	1	1
m								
ɸ	1	3			11		1	
l		5	4	5	8	4	1	
l	7	29	1	6	7	8	1	1
r	16	36	9	5	22	15	5	15
w								
y								

	b	d	g	p	t	k	č	ǰ
n	144	1482	72	118	1343	204	76	155
m	181	63	26	369	80	55	9	9
ŋ	39	37	125	41	77	146	7	10
m		1						
ɸ	8	28		1	141	2	7	
l	24	75	40	55	84	47	10	1
l	71	348	15	74	129	79	14	15
r	156	359	93	86	277	173	41	121
w								
y								

	b	d	g	p	t	k	č	ǰ
b	3	5	1	2	4	2	1	3
d	25	5	3	5	14	8	2	3
g	4	3	3	2	3	1	1	1
p	4	1	2	4	11	2	2	1
t	18	14	6	11	18	13	1	3
k	2	6	3	7	45	6	1	1
č	2	2	3	1	4	3	2	1
ǰ	1	8	1	3	4	1	2	2
s	9	7	4	46	158	56	7	1
z	15	21	9	11	16	17	2	4
ʃ	1	1	1	1	4	1	1	
ʒ		1						
f	1	1	1	2	12	3	1	
v	1	3	1	7	3	5	1	1
θ		5		2	1	2		1
ð	2	1		1		1	1	
h								

	b	d	g	p	t	k	č	ǰ
b	24	41	8	14	30	14	7	27
d	283	82	51	95	196	105	16	30
g	28	30	21	17	21	8	7	7
p	33	11	16	32	138	18	15	7
t	210	161	75	119	214	154	15	24
k	26	48	31	58	513	59	26	7
č	16	17	22	9	46	26	14	8
ǰ	12	67	8	25	31	9	14	14
s	98	77	39	503	1778	523	69	12
z	174	221	85	124	189	175	19	40
ʃ	8	8	7	8	44	7	7	
ʒ		7						
f	13	17	8	15	155	31	7	
v	36	77	21	67	50	61	8	9
θ	7	38	3	16	14	24		7
ð	14	8		7		7	7	
h								

## MIT

	n	m	η	μ	ρ	l	l	r	w	y
p	23	15				1	15	11	15	2
e	29	11					8	1	2	1
o	29	1	10				39	61		
o'	24	6					20	77	3	
u	15	23				1	9	5	7	5
o'	16	15				1	15	12	2	2
o'	9	1					4			
o'	20	2					2	10	2	
z	8	7					4	1	5	
i	112	32	46				49	24	2	
e	56	17	3				48	26		
æ	125	24	8				23	31		
a	17	15	1				16	100		
Λ	46	34	11				11			
U		2					16	30		
ø	7	5				5	6	2	13	2
ə	104	62					50	10	10	4
i	94	1	60							

## APDB

	n	m	η	μ	ρ	l	l	r	w	y
p	276	224				41	213	143	219	31
e	293	157				6	109	32	43	10
o	335	10	127				480	574		2
o'	293	92				1	258	987	38	3
u	161	247				60	115	72	78	57
o'	220	171				14	161	151	22	17
o'	83	7				5	55		1	2
o'	241	22				2	21	122	20	2
z	104	79				6	63	19	43	6
i	1372	423	500				509	330	16	
e	795	213	27				540	340		
æ	1544	233	90			1	231	350		
a	201	160	8				183	1022		
Λ	522	437	108				121			
U		28					163	299		
ø	74	58				63	62	35	110	20
ə	1324	680	6				537	162	152	37
i	1013	37	758				5			

	n	m	η	μ	ρ	l	l	r	w	y
n	5	15				8	9	3	8	8
m	3	10				2	5	2	3	10
η	2	1					2	5	5	4
μ							1			
ρ						2			1	
l	2	1					4	3	2	
l	2	8					3	3	12	8
r	16	23			1	2	18	5	13	6
w						1				
y					1					

	n	m	η	μ	ρ	l	l	r	w	y
n	59	169				88	129	50	113	106
m	33	88				33	49	24	53	90
η	23	28					33	48	50	31
μ	1						7		2	
ρ	1	1				20	7	1	9	
l	24	20					85	31	24	2
l	27	85			2		29	41	115	81
r	160	262			9	34	167	55	142	58
w						8		1		
y						7	1			

	n	m	η	μ	ρ	l	l	r	w	y
b	1	1				21	21	27	2	9
d	8	5			11	4	7	32	10	5
g	3	2				4	8	36	6	5
p	4	3				10	41	57	4	10
t	6	10		5	5	7	18	63	16	14
k	2	3			1	25	36	30	26	18
c	1	1			1	1	2	1	1	2
j	1	2					1	1	1	1
s	5	18			6	2	16	4	18	5
z	9	14		2	7		4	9	14	7
z	1	3			7	5	2	2		1
z		1								
f	1	1				6	12	35	1	6
v	8	5			1	6	3	12	3	5
θ	1	1					1	12	1	
δ				1						1
h								18	5	

	n	m	η	μ	ρ	l	l	r	w	y
b	8	9				236	228	274	17	82
d	114	106			116	43	123	330	146	55
g	46	19				43	98	357	47	56
p	30	33				102	425	633	35	85
t	99	152		39	60	102	222	728	250	125
k	36	37			7	251	376	324	287	185
c	9	12			7	15	20	10	13	14
j	8	16				4	10	7	9	11
s	66	184			70	31	197	54	213	47
z	154	163		29	100	8	71	98	175	74
z	10	24			49	55	17	22	1	7
z		7							1	
f	10	11				98	149	330	16	71
v	68	71			8	74	44	129	37	54
θ	9	14			1	1	16	128	15	1
δ		1		7						7
h						1			315	44

## MIT

	s	z	ʒ	ʒ	f	v	θ	ð	h
p	28	43	3	1	11	18	2	6	6
c	14	12	20	2	2	5	1		1
ɔ	9	6	2		12		5		2
o	14	18	5	2	2	16	4	1	2
u	18	25	3		8	9	5	10	8
ɑ	15	17	1		2	8		4	2
ɔ	7	5						1	
ɑ	6	1							
ɜ	10	7	1		5	6	4	2	1
i	50	64	32	2	21	22	7	24	
e	27	3	7	2	4	25		3	
æ	17	15	6	1	17	15	3	1	
ɑ	11	1	2	5	2	2			
ʌ	16	2	3		4	7	2	7	
u									
ə	7	36	2		9	4	1	7	8
ɔ	67	54	3		28	55	6		14
i	43	17			2	2	1		

## APDB

	s	z	ʒ	ʒ	f	v	θ	ð	h
p	374	435	52	9	170	194	56	102	158
c	194	121	248	15	36	72	13	10	22
ɔ	106	55	19		147	1	45	5	14
o	174	184	59	17	24	165	51	29	33
u	193	216	38	19	88	91	53	131	109
ɑ	154	211	11		53	94	3	42	23
ɔ	67	45			2			7	
ɑ	55	22	2		2	5	7	7	7
ɜ	129	97	17		51	68	44	23	13
i	647	901	290	21	252	251	152	168	5
e	332	43	67	23	47	286	12	33	1
æ	206	263	67	9	166	180	28	14	1
ɑ	119	11	16	36	18	19	2	8	1
ʌ	238	52	30		58	83	25	111	
u		2	5		1				
ə	100	346	23		86	40	10	83	95
ɔ	797	639	66		325	825	66	2	174
i	452	162	12		34	46	7		1

	s	z	ʒ	ʒ	f	v	θ	ð	h
n	60	45	7		12	1	3	34	11
m	6	20	1		8	1	2	11	4
ŋ	10	5	2		5	2	2	4	8
m		2							
ɔ	2	4			1	1		1	
l	6	17			6	1	1	1	
l	7	13	1		11	3	5	5	4
r	15	16	2		8	3	2	9	1
w									
y									

	s	z	ʒ	ʒ	f	v	θ	ð	h
n	718	451	83		165	43	37	367	180
m	93	213	9		81	9	20	97	58
ŋ	94	61	17		48	18	23	54	62
m		18	1					1	
ɔ	25	36			9	7	1	10	2
l	64	169	4		67	12	10	11	24
l	123	152	16		140	52	39	57	58
r	191	157	25	1	101	30	27	110	60
w									5
y									

	s	z	ʒ	ʒ	f	v	θ	ð	h
b	2	4	1		1	1	1	1	1
d	15	17	1		15	3	4	15	11
g	1	11	1	1	1		1	1	1
p	13	1			2	1	1	3	1
t	77	1	2		12	2	1	21	12
k	53	1	4		6	1	1	4	2
j	1		1		1		1	2	1
s	7	1	7		10	1	2	7	7
z	19		9		19	2	2	5	10
ʒ	3	1	1		3	1	1		
ʒ	1								1
f	6				1	1	3	1	
v	4	9	1		6	1		11	5
θ	3	1			2	1	1		
ð	1				2	1		4	1
h									

	s	z	ʒ	ʒ	f	v	θ	ð	h
b	32	36	7		8	12	7	7	8
d	191	233	25		174	54	43	191	214
g	10	131	8	7	10		8	8	10
p	152	7	11		23	7	7	35	14
t	910	7	41		158	24	18	272	233
k	604	7	81		66	11	8	49	47
ç	32	7	7		16	7	7	14	14
j	11		7		9		7	16	11
s	92	7	61		127	14	21	95	107
z	227	1	80		185	26	22	129	204
ʒ	22	7	7		22	7	9	1	3
ʒ	7							1	7
f	62		7		10	7	21	17	14
v	72	98	10		58	13	2	152	71
θ	35	7	2		20	9	8	16	16
ð	7	3			15	7		30	7
h									

## **B Automatic Alignment System**

Leung, H. C., and V. W. Zue, "A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech," *Proc. ICASSP-84*, pp. 2.7.1-2.7.4, March 1984.

# A PROCEDURE FOR AUTOMATIC ALIGNMENT OF PHONETIC TRANSCRIPTIONS WITH CONTINUOUS SPEECH\*

Hong C. Leung  
Victor W. Zue

Department of Electrical Engineering and Computer Science  
Research Laboratory of Electronics  
Massachusetts Institute of Technology  
Cambridge, Massachusetts 02139

## ABSTRACT

A system for automatic alignment of phonetic transcriptions with continuous speech has been developed. The speech signal is first segmented into broad classes using a non-parametric pattern classifier. A knowledge-based dynamic programming algorithm then aligns the broad classes with the phonetic transcriptions. These broad classes provide "islands of reliability" for more detailed segmentation and refinement of boundaries. By doing alignment at the phonetic level, the system can often tolerate inter and intra-speaker variability. The system was evaluated on sixty sentences spoken by three speakers, two male and one female. 93% of the segments are mapped into only one phoneme. 70% of the time the offset between the boundary found by the automatic alignment system and a hand transcriber is less than 10 ms. The performance can be improved by applying more heuristic rules.

## INTRODUCTION

The alignment of a speech signal with its corresponding phonetic transcription is an essential process in speech research, since the time-aligned transcription can serve as pointers to specific phonetic events in the waveform. If a sufficient amount of time-aligned acoustic data is available, speech researchers will then be able to quantify the properties of phonetic segments and describe how their characteristics are modified by contexts. These results in turn will lead to a better model for speech production, as well as better rules for speech synthesis and recognition.

Traditionally, the alignment is done manually by a trained acoustic phonetician, who listens to the speech signal and visually examines various displays of the signal. There are several disadvantages to this approach. First, the task is extremely time consuming; even under the best of circumstances, the process of time alignment can take several minutes for one second of speech material. Second, the task requires the skill and knowledge possessed by a small number of experts. These two reasons combine to severely limit the amount of data that can be collected in this manner. Third, there is the lack of consistency and reproducibility of the results. Manual labeling often involves decisions that are highly subjective. Even if the sentence and the transcription were the same, the inter- and intra-transcriber variability can still be quite high. Finally, there is the problem of human error associated with tedious tasks.

The problems associated with manual labeling, together with the need for a large corpus of time-aligned data, clearly call for the development of an automatic time-alignment system. From the

practical standpoint of developing a phonetically-based speech recognition system, automatic time alignment will not only help enhance our basic acoustic-phonetic knowledge, but also provide a testbed for specific recognition algorithms. In other words, knowing what the phonetic strings are should make it easier for us to find the phonetic segments.

Over the past few years, several automatic time alignment procedures have been suggested in the literature. Most of these approaches attempt to align the speech waveform with a reference waveform, using dynamic programming algorithms. The reference waveform may be a known and previously labeled utterance [3,4], a concatenation of stored templates [5], or a synthetically generated utterance [6]. In order for these methods to be effective, the two waveforms must not differ significantly in detailed phonetic structures, or the synthesis rules must be fairly advanced. A second approach, which also uses dynamic programming, is to segment and label the waveform into broad phonetic classes prior to time alignment [7]. A more detailed frame-by-frame labeling is then achieved by a second dynamic programming algorithm, using derivatives of energy and formant functions.

This paper describes a new method of automatic phonetic alignment. This method utilizes a standard pattern classification algorithm, a dynamic programming algorithm, and the constraints imposed by our acoustic-phonetic knowledge. The speech signal is first segmented into broad phonetic classes using a non-parametric pattern classifier. The resulting string is then aligned with the transcription using a knowledge-based dynamic programming algorithm. Acoustic phonetic knowledge is utilized extensively in the feature extraction for pattern classification, the specification of constraints for time-aligned paths, and the subsequent segmentation/labeling and refinement of boundaries.

## SYSTEM DESCRIPTION

The basic structure of the system that we have developed is shown in Figure 1. The speech signal is digitized at 16 kHz and captured by an automatic end-point detection algorithm [1]. From the speech signal, a number of parameters are computed once every 5 ms. These parameters are then used in conjunction with a pattern classifier to produce 6 broad phonetic classes. The output of the classifier is used to time-align major and robust phonetic events with the transcription in a way similar to that proposed by Wagner [7]. This initial time alignment serves as anchor points for subsequent detailed phonetic alignment utilizing a set of heuristic rules.

### Initial Broad Classification

Ideally, one would like to directly classify the speech signal into segments that correspond to detailed phonetic events. However, the

\* Research supported by the Office of Naval Research under Contract N00014-82-K-0727 and by the System Development Foundation

task is difficult due to the high degree of acoustic variability in the speech signal. Our approach is to make an initial broad classification relying on traditional statistical pattern techniques. The objective is to determine robust acoustic-phonetic events that are relatively context-independent, and to use these events as anchor points for more detailed analysis. We have chosen to structure the classifier as a sequence of binary classifiers arranged in a binary decision tree. One possible advantage of using a sequence of classifiers is that a different feature vector can be used for each classifier in order to maximize the contrasts between the two possible output classes. For example, zero-crossing rate is helpful for distinguishing sonorants and obstruents, but not for distinguishing vowels from other voiced consonants. Thus the problem of classifying the speech signal into different classes can be reduced to a sequence of sub-problems, which are relatively easier to tackle.

At each node in the decision tree, a binary decision is made by a pattern classification machine as shown in Figure 2. The structure of each of the classifiers is identical; the only difference is the feature vectors and initial seed points used in the clustering algorithm. These classifiers have no knowledge about the transcriptions, make no assumptions about the distributions of the feature parameters and need no training. Each classifier starts with a set of  $M$  parameters selected based on acoustic-phonetic knowledge and computed once every 5 ms. Note that the number of parameters used,  $M$ , may be different for each of the binary classifiers. The parameters are then processed through a seven-point median smoother, clipped, and normalized. Clipping is intended to emphasize the portions of the speech signal where boundaries are likely to occur. Clipping thresholds are selected conservatively such that segment boundaries fall within the transitional regions. Normalization then transforms each of the clipped feature parameters to the same scale. The clipping and normalization procedure effectively assigns different weights to different feature parameters depending on how much the feature parameter distributions of the two classes overlap.

Every 5 ms, an  $M$ -dimensional feature vector is obtained. All the feature vectors for a given sentence constitute samples in the feature space. A binary decision is made in the  $M$ -dimensional feature space by means of K-Means clustering, using a Euclidean distance metric [2]. It is well known that the location of the cluster centers and the speed of convergence for a clustering algorithm depends on the choice of the initial seed points, the number of clusters, and the geometrical distributions of the data. The use of a binary classifier has the advantage that the algorithm is guaranteed to converge. In addition, the binary classifier enables us to apply our acoustic-phonetic knowledge and select initial seed points at the extrema of the feature space to maximize the contrast. We found that the algorithm typically converges after less than 10 iterations.

At the top of the decision tree, the clustering algorithm assigns one of two labels to every frame of data. Each group of data will pass through a different classifier at a lower node, and the process repeats. Our experience has shown that the broad phonetic classifier performs very well if the total number of classes is small, say 6 or 7. The performance of the classifier degrades substantially when one attempts to use it to make fine phonetic distinctions. In our implementation, the classifier assigns one of six labels to every frame of the data: S (vowel-like sonorant), O (obstruent), Vo (voiced-obstruent), Si (silence), B (nasals and voice-bars), and UI (unlabeled segments). UI is assigned to segments with clear evidence of energy dip in the vowel-like sonorant regions. A context-dependent median smoother is used to remove spurious segments, although it is rarely needed.

Figure 2 shows the spectrogram and waveform of the sentence, "A

tusk is used to make costly gifts", spoken by a male speaker. The output of the broad phonetic classifier is shown in row (a). The vertical lines drawn in the spectrogram are at segment boundaries determined by the classifier.

#### Alignment

The output of the initial classifier is a broad, but presumably robust, description of the significant acoustic phonetic events in the speech signal. In order to use this broad phonetic description as anchor points for more detailed analyses, the broad representation must now be aligned with the phonetic transcription. This is essentially a path searching problem, and we have chosen to use dynamic programming, where the path is heavily constrained by acoustic phonetic rules. Figure 4 illustrates how this is done for the same utterance as shown previously. The horizontal dimension represents the output of the classifier, while the vertical dimension represents the actual transcription. (Durational information is used by the algorithm, but is not explicitly represented in this figure. Two kinds of constraints direct the algorithm to search for the correct path. First, the path is not allowed to traverse through certain cells, since this will produce unpalatable phonetic alignments. These mismatches are stored as a set of context-independent rules, and the resulting cells are marked in the figure by an  $x$ . For example, the first phoneme /t/ is not allowed to match a silence or a sonorant segment. Second, there is a set of rules that eliminates certain matches based on contextual information. These cells are marked in the figure with an open circle. For example, the first /t/ is not allowed to match the second obstruent due to a durational constraint. The filled circles denote the optimum path, subject to a predetermined set of cost functions. As can be seen from this example, the acoustic-phonetic constraints can often reduce the number of possible paths dramatically.

Row (b) of Figure 3 shows the results of the time alignment. This sentence contains an example where two segments were aligned with one phoneme at time  $t = 1.85$  sec. It also contains three examples where two phonemes were aligned with one segment at  $t = 1.0$  sec,  $t = 1.45$  sec and  $t = 1.55$  sec. In these latter cases, further segmentation is clearly needed.

#### Knowledge-based Segmentation

The dynamic programming algorithm described previously divides the speech signal into a sequence of segments. Each segment is mapped into one or more phonetic symbols or events. No further processing is needed if the matching is one or more segments to one phoneme. For those segments which correspond to 2 or more phonetic events, further segmentation is achieved by applying a set of heuristic rules. The transition between some phonetic events is gradual and is not marked by any distinct acoustic cues. In these cases, we have chosen to mark the boundary by using a set of ad hoc, but consistent rules. For example, pre- and post-vocalic liquids next to certain vowels are assumed to have a duration that constitutes one-third of the syllable nucleus. Row (c) of Figure 3 shows the results after the application of such ad hoc rules. Note that the pre-vocalic /l/ at time  $t = 1.55$  sec. has been delineated from the following vowel.

The transitions between some phonetic events are pronounced and are marked by clear acoustic cues. In these cases, further segmentation is accomplished by a proper selection of feature parameters and algorithms based on contextual information. Row (d) in Figure 3 contains two examples, at time  $t = 1$  sec. and  $t = 1.45$  sec. The output of the feature-based segmentation compares favorably with the hand transcription shown in row (e).

Some segments are mapped into multiple phonemes. Two examples



are shown in Figure 5. We are continuously adding and refining the phonetic rules. With the proper rules and features, it is hoped that these problems will ultimately be solved.

### EVALUATION

The automatic transcription alignment system described in the previous section was evaluated using sixty sentences spoken by three speakers, two male and one female, representing approximately 150 sec. of speech. The sentences were selected from the Harvard list of phonetically balanced sentences. All three speakers read the same twenty sentences. All sentences were transcribed and manually aligned by an experienced transcriber. There are approximately 1800 phonetic events in the transcription. For comparison, five of the sixty sentences, selected at random, were manually labeled by a second transcriber. The entire process of manual labeling took upwards of 15 hours.

Figure 6 summarizes the number of phonetic events that matches to one segment after two different stages of processing. Approximately 80% of the time there is a one-to-one correspondence after time alignment, whereas the final results is over 90%. In other words, only 7 % of the segments requires further segmentation. This segmentation can presumably be accomplished as more rules and features are used. On the other hand, they can also be corrected by manual intervention.

Figure 7 (a) shows the cumulative distribution of the boundary offsets between the automatic alignment system and the first transcriber for the sixty sentences. Approximately 75% of the time this offset is less than 10 msec. Figure 7 (b) shows the boundary offsets between the two transcribers for five of the sixty sentences. In this case approximately 80% of the time the offset is less than 10 msec.

### SUMMARY

In this paper we described a system that automatically aligns a phonetic transcription with the corresponding speech waveform. The

system performs initial classification by a pattern classification algorithm. The output of the classifier is used to determine "islands of reliability" for further segmentation. Acoustic phonetic knowledge is used extensively during classification, time-alignment, and feature-based segmentation. We are encouraged by the preliminary results, and are hopeful that this system will play a major role in establishing a large database for speech research.

### REFERENCES

- [1] L.F. Lamel et al., "An Improved Endpoint Detector for Isolated Word Recognition," IEEE Trans. Acoustic, Speech, Signal Processing, Vol. ASSP-29, No.4, August 1981.
- [2] J.T. Tou and R.C. Gonzalez, Pattern Recognition Principles, Reading, MA: Addison-Wesley, 1974.
- [3] R.M. Chamberlain, J.S. Bridle, "Zip: A Dynamic Algorithm For Time-aligning Two Indefinitely Long Sequences," in Proc. ICASSP, Apr.1983, pp.816-819.
- [4] H.D.Hohne et al., "On Temporal Alignment of Sentences of Natural and Synthetic Speech," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-31, pp.807-813, August 1983.
- [5] M.R. Lowry, "Automatic Labelling of Speech from the Phonetic Transcription," S.M. thesis, Massachusetts Institute of Technology, 1978.
- [6] M.Lennig, "Automatic Alignment of Natural Speech with a corresponding transcription," Speech Communication, Vol.2, 11th International Congress on Acoustics Satellite Symposium on "The Processes for Phonetic Coding and Decoding of Speech", Toulouse, 15-16 July 1983.
- [7] M. Wagner, "Automatic Labelling of Continuous Speech with a Given Phonetic Transcription Using Dynamic Programming Algorithms," in Proc. ICASSP, Apr. 1981, pp.1156-1159.

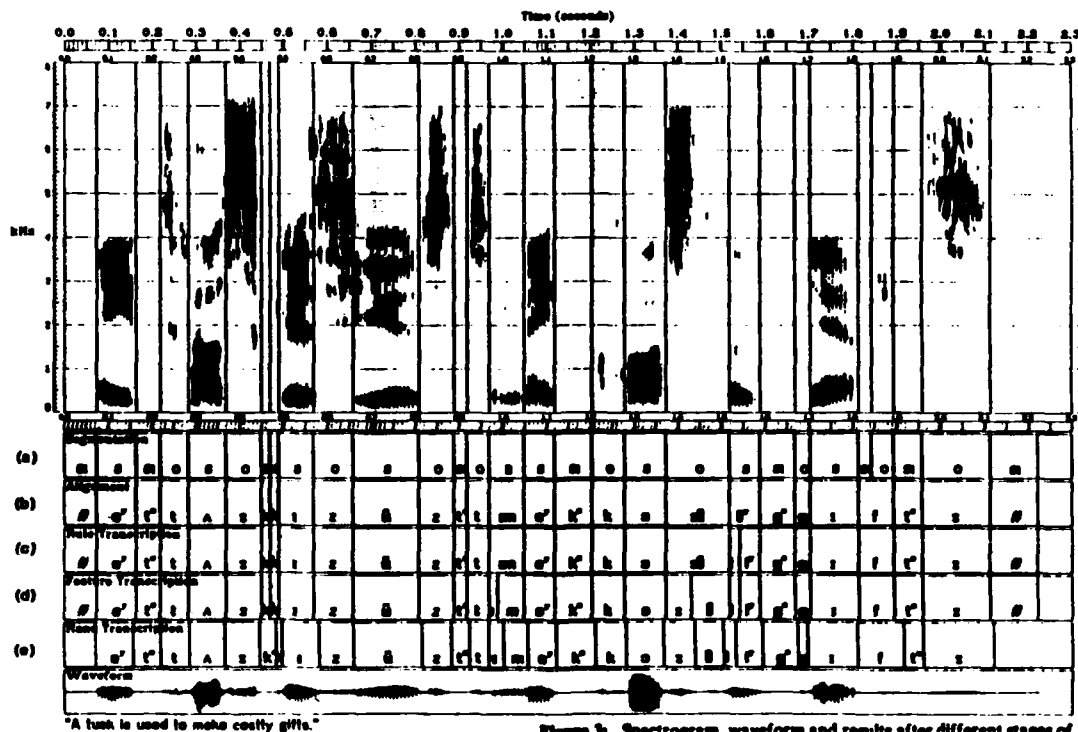


Figure 3: Spectrogram, waveform and results after different stages of processing of a sentence spoken by a male speaker

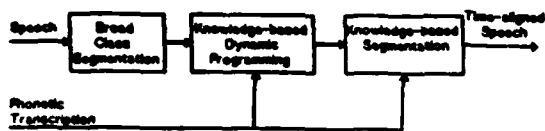


Figure 1: Basic system structure.

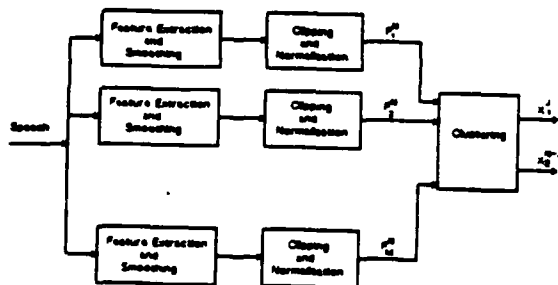


Figure 2: Block diagram of the pattern classification machine. Superscripts denote number of samples.

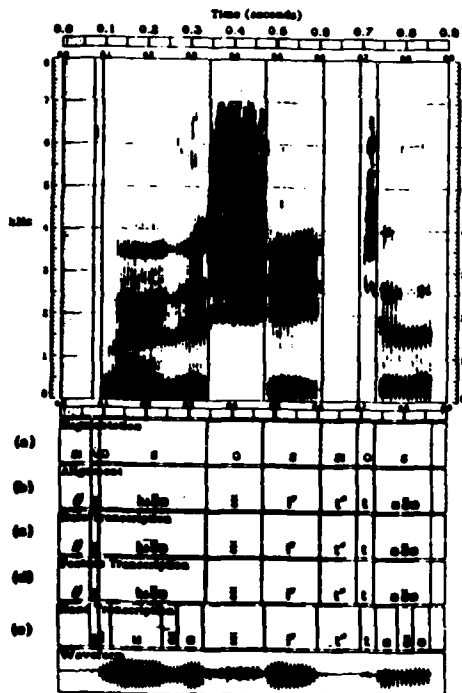


Figure 3: An example for the phrase "Close the sheet to the". Further segmentation is needed.

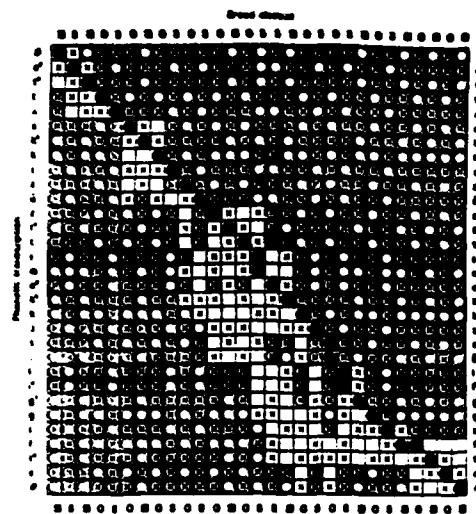


Figure 4: An example of the knowledge-based dynamic programming

Number of phonemes in 1 segment	Knowledge-based Dynamic Programming	Knowledge-based Segmentation
1	79%	93%
2	15%	3%
3	3%	2%
4	2%	1%
5 or more	2%	1%

Figure 6: Statistics on number of phonetic events in 1 segment after two different stages of processing.

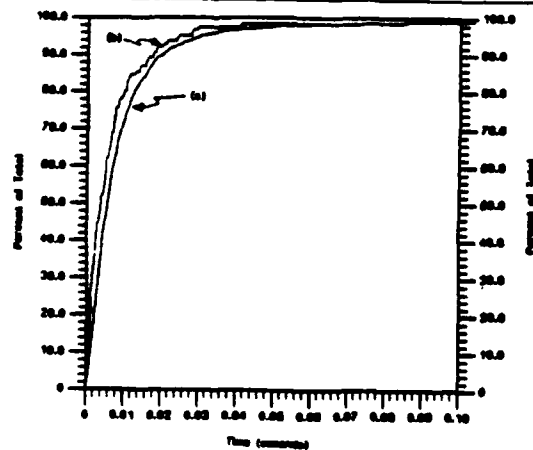


Figure 7: Cumulative distributions of the boundary offset.

## **C Transcription and Analysis**

Zue, V.W., and S. Seneff, "Transcription and Alignment of the TIMIT Database," *Proc. of the Second US/Japan Joint Symposium on Spoken Language Systems*, November, 1988.

## TRANSCRIPTION AND ALIGNMENT OF THE TIMIT DATABASE

Victor W. Zue and Stephanie Seneff

Spoken Language Systems Group  
Laboratory for Computer Science  
Massachusetts Institute of Technology  
Cambridge, MA 01890  
U.S.A.

### ABSTRACT

The TIMIT acoustic-phonetic database was designed jointly by researchers at MIT, TI, and SRI. It was intended to provide a rich collection of acoustic phonetic and phonological data, to be used for basic research as well as the development and evaluation of speech recognition systems. The database consists of a total of 6,300 sentences from 630 speakers, representing over 5 hours of speech material, and was recorded by researchers at TI. This paper describes the transcription and alignment of the TIMIT database, which was performed at MIT.

### 1 BACKGROUND

When the DARPA Strategic Computing speech program was first formulated in 1984, the consensus of the research community was that the amount of speech data available is woefully inadequate. As a result, a significant effort on database development was mounted in order to provide the research community with a large body of acoustic data for research, system development, and performance evaluation. One such database is the so-called TIMIT acoustic-phonetic database. The TIMIT database was designed jointly by researchers at MIT, TI, and SRI. It consists of a total of 6,300 sentences from 630 speakers, representing over 5 hours of speech material, and was recorded by researchers at TI. This paper describes the transcription and alignment of the TIMIT database, which was performed by researchers at MIT.

Each speaker in the TIMIT database recorded 10 sentences drawn from three different corpora as follows. Each speaker read two sentences, designated as S1 and S2, which were designed by Jared Bernstein of SRI in order to compare dialectal and phonological variations across speakers. Five sentences, designated as SX sentences, were drawn from a small set of sentences designed at MIT. The remaining three sentences for each speaker, designated as SI sentences, were selected from the Brown corpus by Bill Fisher of TI [1].

There are a total of 450 "MIT" sentences used in the TIMIT database. These were generated by hand in an iterative fashion, with the goal that they should be phonetically rich. Care was taken to have as complete a coverage of left- and right- context for each phone as possible. Some of the more problematic sequences, such as vowel-vowel and stop-stop, were particularly emphasized. An attempt was also made to ensure that many of the frequently-occurring low-level phonological rules were adequately represented. To aid in the sentence generation process, we made use of an on-line, Webster's Pocket Dictionary containing nearly 20,000 words. Words or word-sequences containing particular phone pairs could be accessed

from this dictionary automatically, which greatly facilitated the database design process. We performed a detailed analysis of the resulting sentence set, as well as the SI sentences that make up the remainder of the database. The interested reader should consult Lamel et al. [3] for further information about the corpora.

## 2 THE ACOUSTIC PHONETIC LABEL SET

All of the recorded sentences were provided with a time-aligned sequence of acoustic-phonetic labels. The label set is intended to represent a level somewhat intermediate between phonemic and acoustic. Our motivation was that clear acoustic boundaries in the waveform should all be marked, and that the criteria for positioning the boundaries between units should in part be based on our ability to mark them consistently. Table 1 lists all of the acoustic-phonetic labels that were used. Most of these labels are phonemic, although several symbols have been included for labelling acoustically distinct allophones as well as other special acoustic events.

### 2.1 Stops

Stops are characterized by a sequence of two events: a closure and a release. This departure from phonemic form is, we believe, important in order to preserve a boundary marking the onset of the release. There are six closure symbols for the stops. The closure region for affricates is identical with that of the corresponding alveolar stop (e.g., the /tʃ/ in "chat" is represented as [tʃ̚]).

There are two major allophones for the stops. The glottal stop, [ʔ], is often inserted preceding a word-initial vowel. Sometimes a /t/ can also be realized as a glottal stop, as in "cotton". The symbol [ɾ] is used to label a flap, which can be either an underlying /t/ or /d/. We make a separate flapping decision for every phonemic /t/ and /d/, based on listening and the spectrographic evidence. We allow flapping to occur in environments for which theory is violated, if in fact we believe that flap is what was heard/seen.

### 2.2 Nasals and Semivowels

We recognize four allophones for the nasals, three of them are the syllabics, [m̩, n̩, ŋ̩]. If there is any evidence of a preceding schwa, the non-syllabic form is preferred. The alveolar nasal, /n/, can be realized as a nasal flap, denoted by the symbol [ɾ̃]. Sometimes an underlying /nt/ sequence is realized as a nasal flap, as in "entertain."

The liquid, /l/, has a syllabic allophone, denoted as [l̩]. Again, a non-syllabic form is preferred whenever a preceding schwa is observed.

### 2.3 Vowels

Two vowels, /i, o/, are represented by symbols that included their corresponding off-glides. This is because they are usually realized as diphthongs in American English. The four diphthongs, /aɪ/, /aʊ/, /ɔɪ/, and /eɪ/, are each represented as a single label, with no separate region defined for the off-glide portion. The retroflexed vowel /ɜ̣/ is also represented as a single unit. This represents a departure from the International Phonetic Alphabet, which would represent this steady-state vowel as the sequence /ʌɾ/.

Reduced vowels are represented by four separate allophones: back schwa ([ɐ]), front schwa ([ɪ]), retroflexed schwa ([ɜ̣]), and voiceless schwa ([ɰ]). The decision for [ɐ] vs [ɪ] is based on whether the second formant is closer to the first or to the third. A low third formant leads

Phonetic Symbol Mapping					
IPA	Char	Notes	IPA	Char	Notes
Stops					
p	P		b	b	
t	t		d	d	
k	k		g	g	
p <sup>h</sup>	⊕	Symbol-+	t <sup>h</sup>	δ	Symbol-shift-D
t <sup>h</sup>	∞	Symbol-i	d <sup>h</sup>	↑	Symbol-g
k <sup>h</sup>	θ	Symbol-p	g <sup>h</sup>	±	Symbol-:
ʔ	F		ʔ	?	
Nasals					
m	m		ɱ	M	
n	n		ɳ	N	
ɳ	G		ɳ	π	Symbol-shift-P
ɱ	ε	Symbol-shift-E			
Fricatives					
s	s		ʃ	S	
z	z		ʒ	Z	
ç	C		j	J	
θ	T		ð	D	
f	f		v	V	
Liquids, Glides, Silence, and h					
l	l		ɭ	L	
r	r		w	w	
ɻ	y				
ɭ	λ	Symbol-shift-L	ɭ	C	Symbol-t
h	h		ɦ	H	
Vowels					
ε	E		ɪ	I	
ɔ	c		æ	⊙	
a	a		ʌ	.	
u	u		ʊ	U	
ɜ	R		ü	:	
a <sup>y</sup>	Y		ɔ <sup>y</sup>	O	
e <sup>y</sup>	e		i <sup>y</sup>	i	
a <sup>w</sup>	W		o <sup>w</sup>	o	
ə	x		ə	X	
ɪ			ɪ	γ	Symbol-shift-G

Table 1: A list of the acoustic phonetic symbols used for the transcription of the TIMIT database

to /ə/. Schwas can often be devoiced in words such as "secure".

English does not distinguish phonemically between the fronted vowel / $\bar{u}$ / and the standard back /u/; however the difference in  $F_2$  for the two forms can be as much as 800 Hz. We felt it was unsatisfactory to group two forms with such diverse formant frequencies into the same vowel category. The decision is made as for schwa: if  $F_2$  is closer to  $F_1$ , it's considered a back /u/. Similar trends of fronting are also observed for /o/ and / $\bar{u}$ / in certain environments; however, the effect is most dramatic for /u/.

At present, we make no attempt to provide further sub-phonemic characterizations for vowels other than this front/back distinction for /u/ and the four schwas. For instance, many vowels are nasalized when they are followed by a nasal, or lateralized when followed by an /l/. Such information would surely be useful, but the decision-making process is prone to judgment error, and would require a significant increase in time and effort.

## 2.4 Others

We make a distinction between two types of /h/: voiced ([ $\bar{h}$ ]) and unvoiced ([h]). The decision is based mainly on an examination of the waveform for clear low-frequency periodicity, and spectrogram for voicing striations. The voiced form is most common between two vowels.

Our label set includes a category "epenthetic silence,"  $\bar{\bar{\bar{u}}}$ , which we use to mark acoustically distinct regions of weak energy separating sounds that involve a change in voicing. These short gaps are typically due to articulatory timing errors. The most common occurrences of such gaps are between an /s/ and a semivowel or nasal, as in "small," "swift," or "prince." Two other non-phonetic symbols are included: # is used to mark regions preceding and following a sentence, and  $\square$  is used to mark pauses within a sentence.

## 3 CRITERIA FOR BOUNDARY ASSIGNMENTS

The acoustic-phonetic transcription for the TIMIT sentences is time aligned with the speech waveform. The alignment is useful in that specific acoustic events can be accessed conveniently based on the transcription. We must stress, however, that the aligned transcription is intended to establish a correspondence between the transcription and important acoustic landmarks. One should not directly associate a region between two time markers as a distinct phonetic unit, since the encoding of phonetic information in the speech signal is extremely complicated.

In most cases, the boundaries between two acoustic-phonetic events are clear and well-defined, such as that between a stop closure and its release. However, there are a number of cases where the exact placement of a boundary is problematic (as is the case between a semivowel and a vowel), or cases where it's not clear whether a region should be represented as one or two acoustic-phonetic units (as is the case for diphthongs). In these cases, we tried to define a set of criteria that would be systematic and least subject to human error, in order to produce boundary positionings that were as consistent as possible.

As mentioned previously, we decided that the boundary between the closure interval and the release of a stop is an important one that should be assigned. It is certainly a very distinct landmark in the waveform. Anyone interested in studying the burst characteristics of a stop would then be able to focus on just that region that includes only the released portion. In a strictly phonemic representation, the closure and release would be represented as a single unit, and therefore that critical boundary would remain unmarked.

A problematic boundary is one that separates a prevocalic stop from a following semivowel, as in "truck." Typically, part of the /r/ is devoiced, and therefore is absorbed into the aspiration portion of the stop. If listening were the only criterion, then the left boundary of the /r/ would occur somewhere in the aspiration, and the right boundary would occur somewhere after voicing onset. A clear acoustic boundary at the point of voice onset would remain unmarked. It would also be difficult to decide where to mark the boundary between the stop burst and the aspirated /r/ portion. Since voice-onset time (VOT) is a parameter that has been a focus of many research efforts, it seems unsatisfactory not to include a reliable mechanism for measuring VOT based on the labelled boundaries. Therefore, we adopted the policy of always absorbing into the stop release all of the unvoiced portion of a following vowel or semivowel.

The boundary between many semivowels and their adjacent vowels is rather ill-defined in the waveform and spectrogram, because transitions are slow and continuous. It is not possible to define a single point in time that separates the vowel from the semivowel. In such cases, we decided to adopt a simple heuristic rule, in which one-third of the vocalic region is assigned to the semivowel, thus giving the vowel twice the duration of the adjacent semivowel. Previous investigators have also made use of such consistent rules for defining acoustically ambiguous boundaries [4].

One obscure condition is a /ts/ or /dz/ sequence, where typically there is little or no spectral change to characterize a boundary between the homorganic stop and fricative, yet the onset of acoustic energy of the unit is sufficiently abrupt such that a /t/ is heard. Our convention here is that, if a clear /t/ is heard, the early portion of the /s/ is marked as a /t/ release.

When gemination occurs, we do not attempt to mark a boundary between the two units. This situation occurs exclusively at word boundaries, as in "some money." Furthermore, in the case of a stop-stop sequence where the first stop is unreleased, the closure interval is assigned to the first stop and the release to the second one.

## 4 PROCEDURE FOR TRANSCRIPTION AND ALIGNMENT

The transcription and alignment process involves three stages:

1. An acoustic-phonetic sequence is entered manually by a transcriber as a string.
2. The speech waveform is aligned automatically with the acoustic-phonetic sequence, using an alignment program developed at MIT.
3. The boundaries generated automatically are then hand corrected by experienced acoustic phoneticians.

### 4.1 Transcription

In both stages 1 and 3, the labeller makes her/his acoustic-phonetic decision based on careful listening of portions of the speech waveform, as well as visual examination using displays such as the spectrogram and the original waveform. The process takes place within the SPIRE software facility for speech analysis, a powerful interactive tool that is well-matched to this task [2]. Stage 1 requires less intensive use of SPIRE than stage 3, because it is only necessary to record what was heard, without identifying the time locations of the events. Furthermore, minor errors of judgment made at this stage can be readily corrected





class labels: *sonorant*, *obstruent*, *voiced-consonant*, *nasal/voicebar*, and *silence*, using a non-parametric pattern classifier. The assignment process makes use of a binary decision tree, based on a set of acoustically motivated features. Each sequence of identically-labelled frames is then collapsed into a segment of the same label, thus establishing a broad-class segmentation of the speech. The output of the initial classifier is then aligned with the phonetic transcription using a search strategy with some look-ahead capability, guided by a few acoustic phonetic rules. For those segments which correspond to two or more phonetic events after preliminary alignment, further segmentation is done using specific algorithms based on knowledge of the phonetic context. In some cases heuristic rules are invoked (as between a vowel and a semivowel) to assign consistent, but somewhat arbitrary boundaries.

Over the past two years, two major modifications of CASPAR have taken place. First, the alignment of the broad-class acoustic labels with the phonetic symbols has been cast into a probabilistic framework. By using a large body of training data, a set of robust, context-dependent and durational statistics were obtained. Second, a fourth module has been added to the system to improve the resolution of the boundaries. This module computes appropriate acoustic attributes at a high analysis rate using different window shapes that depend on the specific context. The boundaries are then adjusted based on these attributes.

In a formal evaluation, it was found that CASPAR can correctly perform over 95% of the labeling task previously done by human transcribers. The boundary locations produced by the system agree well with those produced by human transcribers. For example, over 75% of the automatically generated boundaries were within 10 msec of a boundary entered by a trained phonetician.

Figure 2 displays the output for the sentence, "She had your dark suit in greasy wash water all year." The transcription and boundaries are overlaid on the spectrogram for ease of examination. For this example, most of the boundaries have been found correctly by CASPAR. Note, however, that boundaries are missing in the [ihæ] sequence of "She had." The waveform displays the word "dark" and the [s] of "suit." Note that the initial boundary of the first [d] is slightly too far forward in time.

### 4.3 Post-Processing

The final step is to correct by hand any errors in the automatically aligned acoustic-phonetic sequence. Some of the errors are due to the fact that CASPAR is not able to determine certain boundaries, such as some of those between two vowels. In other cases the boundaries may have been misplaced.

Hand correction of the aligned transcription is based on critical listening of portions of the utterance as well as visual examination of the spectrogram and the waveform. The spectrogram covers close to 3 seconds worth of speech at one time, whereas the waveform is displayed on a much more expanded time scale. For example, to accurately mark the onset of the release of a stop, the cursor is first positioned on the spectrogram at the approximate point in time. The waveform display automatically moves to synchronize in time with the cursor, and a fine-tuning of the boundary can be achieved by mousing the exact time point in the waveform.

The mouse can be used with ease to move an existing boundary to a new point in time, to erase a boundary, or to insert a boundary. Furthermore, a specified mouse click on any segment allows the labeller to change the acoustic-phonetic label associated with that segment. This step is sometimes necessary to correct an error of judgment made in stage 1.

An example of the screen layout used for the correction process is shown in Figure 3. The



# Phonetic Transcription Layout 1

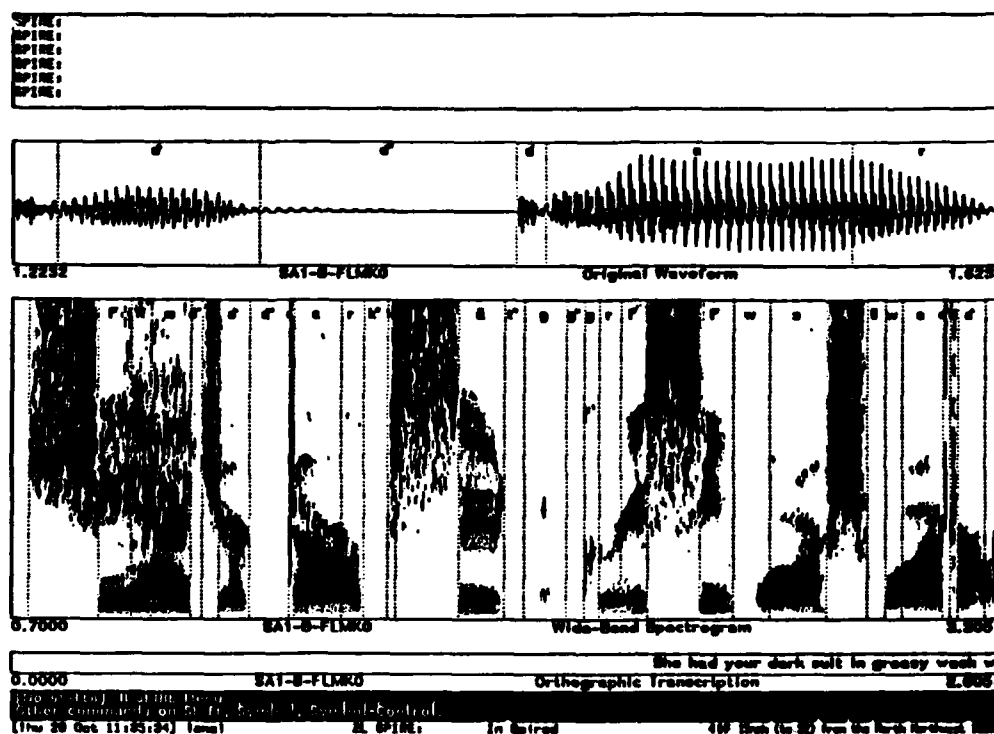


Figure 3: SPIRE layout showing the aligned transcription following post-processing

other transcriptions in future releases.

The transcription and alignment of the TIMIT database is a sizable project. At this writing, all of the sentences have been processed and delivered to the National Bureau of Standards. A significant portion of the database is now available to the general public via magnetic tapes, and plans for distributing them by way of compact disc is well under way. Despite our best intention to provide as correct a set of transcriptions as possible, however, errors undoubtedly exists. We urge users of this database to communicate errors to us whenever possible, so that future users can benefit from this effort.

Finally, we would like to thank Dave Pallett, Jim Hieronymus, and their colleagues at NBS for the cooperation, patience, and good humor that they provided. Their help, particularly regarding data transfer, verification, distribution, and fending off eager inquiries, have been indispensable to this project.

The development of the TIMIT database at MIT was supported by the DARPA-ISTO under contract N00039-85-C-0341, as monitored by the Naval Space and Warfare Systems Command. Major participants of the project at MIT include Corine Bickley, Katy Isaacs, Rob Kassel, Lori Lamel, Hong Leung, Stephanie Seneff, Lydia Volaitis, and Victor Zue.

## REFERENCES

- [1] Fisher, W.M. and G.R. Doddington, "The DARPA Speech Recognition Research

Database: Specification and Status," Proceedings of the DARPA Speech Recognition Workshop, Palo Alto, CA, February 19-20, 1986, pp. 93-99.

- [2] Zue, V.W., D. S. Cyphers, R. H. Kassel, D. H. Kaufman, H. C. Leung, M. A. Randolph, S. Seneff, J. E. Unverferth, III, and T. Wilson, "The Development of the MIT LISP-Machine Based Speech Research Workstation," Proceedings of ICASSP-86, Tokyo, Japan, Apr. 8-11, 1986.
- [3] Lamel, L. F., R. H. Kassel, and S. Seneff, "Speech Database Development: Design and Analysis of the Acoustic-phonetic Corpus," Proceedings of the DARPA Speech Recognition Workshop, Palo Alto, CA, February 19-20, 1986, pp. 100-109.
- [4] Peterson, G. and I. Lehiste, "Duration of Syllable Nuclei in English," J. Acoust. Soc. Am., Vol. 32, 693, 1960.
- [5] Leung, H. C. and V. W. Zue, "A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech," *Proc. ICASSP 84*, pp. 2.7.1-2.7.4, March 1984.
- [6] Leung, H. C., "A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech," S.M. Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, January, 1985.
- [7] Kassel, R.H., "Aids for the Design, Acquisition, and Use of Large Speech Databases," S.B. Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, May, 1986.